METHODS OF ESTABLISHING THE RELIABILITY AND VALIDITY OF TESTS

DAVID E. COONEY

Digitized by the Internet Archive
in 2017 with funding from
University of Alberta Libraries

https://archive.org/details/methodsofestabli00davi

THE UNIVERSITY OF ALBERTA

METHODS OF ESTABLISHING THE RELIABILITY AND VALIDITY OF TESTS

A DISSERTATION SUBMITTED
TO THE COMMITTEE ON GRADUATE STUDIES
IN PARTIAL FULFILMENT OF THE DEGREE OF
MASTER OF EDUCATION

FACULTY OF EDUCATION

BY

DAVID EUGENE COONEY

EDMONTON, ALBERTA

April, 1952

ACKNOWLEDGEMENTS

The writer wishes to express his sincere appreciation to Dr. G. M. Dunlop, for his patient encouragement and assistance during the entire preparation of this work, to the sixteen teachers on the Edmonton Public School staff who administered and marked the tests, to the "random sample" pupils who endured another test in order to make the compilation of this data possible, and to a faithful group of my pupils at Garneau Junior High School who willingly assisted in the tabulation of data.

TABLE OF CONTENTS

Chapter		Page
I.	THE PROBLEM	1
II.	THE RELIABILITY AND VALIDITY OF EXAMINATIONS	4
	Reliability	4
	Validity	8
	Item Validity Techniques	10
III.	DESIGN OF THE STUDY	21
IV.	THE RELIABILITY OF THE PICTURE TEST IN SCIENCE	29
	The Split-Half Method	29
	The Method of Rational Equivalence	30
	The Analysis of Variance Method	31
V.	THE VALIDITY OF THE PICTURE TEST IN SCIENCE	39
VI.	THE DAVIS ITEM ANALYSIS TECHNIQUE	42
VII.	OTHER ITEM VALIDITY TECHNIQUES	51
	Flanagan Method	51
	Kelley Method	52
	Standard Error of the Difference	54
	Lawshe Method	56
	Bi-serial r Method	56
VIII.	CONCLUSIONS	73
BIBLIOGRAPHY		77
FOOTNOTES		79
APPENDIX - PICTURE TEST IN SCIENCE		81

LIST OF TABLES

Table		Page
I.	Calculation of Mean and Standard Deviation for 400 Test Scores	24
II.	Tabulation of Choices on Each Item for Item Validity Computation	25
III	Distribution of 400 Test Scores into Odd-Even Classification for Determination of Reliability Coefficient by Split-Half Technique	34
IV.	Summary for Student-Item Chart	37
V.	Analysis of Variance Table for Picture Test in Science with Computed Values	38
VI.	100 Grade VII Test Scores with Corresponding Criterion Scores for Determination of Test Validity	40
VII.	Percentages Suggested for Entering the Davis Item-Analysis Chart when the Percentage of Successes in the Lowest 27% of the Sample is Zero or Negative	47
VIII.	Percentages Suggested for Entering the Davis Item-Analysis Chart when the Percentage of Successes in the Highest 27% of the Sample is Closest to 100	47
IX.	Summary of Discrimination and Difficulty Indices Obtained from Davis Item-Analysis Chart	48
X.	Equivalent Values of Product-Moment r (Flanagan), Fisher's z, and Davis Discrimination Indices	50
XI.	Summary of Discrimination Coefficients Obtained by Flanagan Method	58
XII.	Summary of Data for Kelley Item Validity Method	59
XIII.	Percentages of Area of the Normal Curve	61
XIV.	Summary of Data on Individual Items for Standard Error of the Difference Between Two Percentages	63

LIST OF TABLES - (Concluded)

Table		Page
XV.	Frequency Distribution of D-Values of Items by Lawshe's Nomograph.	68
XVI.	Calculation of Bi-serial r on Item 21.	69
XVII.	Summary of Bi-serial r Calculations on Items 21, 35, 43, 52, and 70	70
XVIII.	Items on the Picture Test in Science which fail to Discriminate as Shown by a Combined Summary of Five Methods.	71

LIST OF FIGURES

Figure — Page

1. Correlation Chart for Determination of Test Reliability by Split-Half Method 36

2. Correlation Chart for Determination of Test Validity 41

3. Chart to Facilitate Use of Kelly Technique 62

4. Lawshe's Nomograph for Assigning Discrimination Values to Test Questions 67

CHAPTER I

THE PROBLEM

Accurate measurement is an indispensable part of every modern activity. Products of industry must be tested for compliance with standards. Careful accounting must be done in business enterprises to determine the efficiency of procedures. All aspects of scientific and engineering projects must be exhaustively checked to ensure the desired performance. In the process of such evaluating, a large variety of precise and dependable instruments of measurement have been constructed.

In a business of the financial magnitude and social importance of education the use of scientific measuring instruments is no less essential. Means have been devised to determine how effectively the objectives of education are being realized in terms of pupil growth and achievement. The task of perfecting the required instruments has been complicated by the subjective nature of many of the goals and the impossibility of keeping the numerous variable factors under control. In spite of the difficulties, however, progress is being made. Errors of measurement are being reduced and better testing methods are being developed.

While the scientific approach to measurement is still comparatively young, educational testing is by no means an entirely new development. The evolving movement can be traced through three stages: oral questioning, written examinations, and the scientific testing movement. Oral questioning, still widely used and still of definite value under specific conditions, was formerly the chief means of testing progress. One of the first scientific advances made in the written or essay type examination which followed can be credited to Horace

Mann[1]. In 1845 he advocated the use of a large number of questions and he commented on the desirability of standardization.

Shortly after 1900 the scientific movement in education was responsible for producing a flood of objective evidence on the unreliability of essay tests. This resulted in a growing awareness that educational testing could and should be approached in a scientific manner. Many forces played a part in the birth and development of the testing movement. The work of Thorndike and his students is usually accepted as the start, although it must be remembered that many previous educators had made important contributions that were necessary to the development of new concepts and techniques. Thorndike is credited in 1904[2] with the first published book dealing directly with measurement. The first standardized tests were in the fields of arithmetic (1908)[3] and handwriting (1910)[4] and before long many others made their appearance. Although a few of the early tests were of excellent quality, the rush to publish and later to commercialize resulted in a flood of tests, hastily constructed, inadequately standardized, and full of technical imperfections.

Standardized tests differ from the ordinary informal objective tests made and used by the teacher in the classroom. The test items used in the two may be similar, but the methods of construction and the subsequent use of the test scores are more refined in the standardized form. In the informal test, the teacher often prepares a number of questions and then gives the test to the pupils, usually as a means of helping to determine a mark on their work. A competent maker of standardized tests proceeds more scientifically. He is more specific in defining what he hopes to measure, but the major difference in procedures is that he tries out the items before they are inserted in the final test. He finds out ahead of time whether each item is satisfactory by giving one or more pretests to a

carefully selected group of pupils. Following this tryout, each item in the test is scrutinized carefully. The final test is composed of items that have proved to be satisfactory. It has been standardized on the basis of the carefully selected group of students. Their scores represent typical achievement levels, called norms, and the scores of any pupils taking the test may be compared with the attainments of the selected group. The scores on a carefully prepared standardized test have more meaning and may be interpreted more objectively than those obtained on an ordinary informal classroom test.

Closely related to the development of the standardized tests was the origin of statistical methods for handling the mass of data accumulated. These new techniques were the means of illustrating and emphasizing the inferiority of many tests that had previously been considered satisfactory.

In this thesis it is proposed to evaluate a number of current methods for measuring the reliability and validity of tests. It is also proposed to compare six of the principal techniques used in determining the validity of test items.

The methods selected for appraisal will be analysed through application to an experimental edition of a picture test in science. The test and the experimental design will be described in Chapter III.

CHAPTER II

THE RELIABILITY AND VALIDITY OF EXAMINATIONS

The evolution of school examinations outlined in the previous chapter has shown that the standardized objective test would seem to be the best testing instrument in use today. Before discussing the statistical procedures used in the standardization of an objective test, it is advisable to stop to examine the two main requisites of a standardized test - reliability and validity.

RELIABILITY

"A test is said to have reliability if it measures consistently what it is supposed to measure."[5] A perfectly reliable test would yield approximately the same results every time it was used. This ideal is often approached, but never, even in physical measurement, completely attained. The test itself often falls short of perfection, but the variations in the pupils measured and the variations in conditions and methods of test administration often contribute most to the variations in test results. A reliable test is relatively free from chance errors of measurement, and scores secured on it are stable and trustworthy. If a pupil makes a score of 75 on a reliable test, we feel confident that this score represents his true ability. If a pupil makes the same score of 75 on an unreliable test, such a score would be subject to large errors of measurement and it would be neither stable nor trustworthy. The results from subsequent testings of an unreliable test will reveal many discrepancies between scores obtained by the same pupils on different occasions.

Experience with tests has contributed rather definite knowledge as to what factors raise or lower test reliability. Symonds[6] compiled a list of factors which influence test reliability.

1. The greater the number of items, the more reliable the test.

2. On the whole, the longer the test time, the greater the reliability. There is, however, an optimal length of time for highest reliability.

3. The narrower the range of difficulty of items, the greater the reliability.

4. Interdependent items tend to reduce the reliability. Such items are passed or failed together, thus reducing the effective length of the test.

5. The more objective the scoring, the greater the reliability.

6. Inaccuracy in scoring lowers reliability.

7. The chance factor in answering test items is important; the greater the probability of success by chance, the lower the reliability.

8. The more homogeneous the material, the greater the reliability.

9. The more common the experiences of the individuals tested, the greater the reliability.

10. Catch questions lower reliability.

11. Various subtle factors which lead to misinterpretation of the item lead to unreliability:

(a) the emotional coloring of words used;

(b) short items in preference to long ones;

(c) choice of words;

(d) poor sentence construction;

(e) inadequate or faulty directions;

(f) misleading intent of the question.

12. Speed of work on a test. Some distribute their time properly; some do not.

13. The mental set for accuracy is important for reliability.

14. Variations in incentive or effort are important.

15. Distractions have some effect, although they may be overrated.

16. Time interval between tests. This is relatively unimportant up to a year for tests of ability, but reliability generally decreases with intervening time.

17. Failure of the individual to grasp the proper mental set and to hold it throughout the test.

Three procedures are used widely for determining test reliability. These are the test-retest method, the equivalent forms method, and the split-half method. In addition to these there are two special methods which will yield reliability coefficients. These are the method of rational equivalence, and a special by-product estimation which can be secured by the analysis of variance technique.

Test-retest method. This is the simplest method of determining test reliability. It consists of the repetition of the test with the same group of subjects, and the correlation of the two sets of scores. There are, however, generous limitations to the dependability of this procedure. If the time between the testings is too short, the subjects will remember many of the answers which they gave on the first trial. This factor tends to make the correlation between the two sets of scores higher than the true reliability coefficient. If speed is important in taking the test, practice on the first trial is very likely to improve performance on the second. This would again make the estimated reliability too high. Memory and practice factors can be partially eliminated by lengthening the interval between the two testings. The danger here is that during the interval some of the subjects will have received training in the function tested or that maturational changes may affect performance. This method is also subject to the

criticism that it is difficult to maintain comparable testing conditions for the two testings, and that it is difficult to sustain motivation throughout the second administration. These difficulties are so serious that the test-retest method is rarely used.

Equivalent forms method. Two tests that are sufficiently similar to be used interchangeably are called equivalent forms of a test. The two forms must be matched carefully for difficulty, content, and organization. The equivalent form is used instead of repeating the same test as in the test-retest procedure, and the reliability is determined from the correlation of results from the two forms. Standard psychological and educational tests for which alternate forms are available make use of this method. The major difficulty with this technique lies in the assumption that the two forms are comparable. Comparability is not a quality to be assumed without evidence, but one to be established by trial. If the two forms of the test are not comparable the correlation between them may be lower than the true reliability of either form.

Split-half method. By this method the test is broken into two equivalent parts and the correlation of the half-tests is then computed. This method may be used when a single form of the test has been administered only once. The method most frequently used for dividing the test into two parts is to take the odd-numbered items for one part and the even-numbered items for the other part. The correlation between the scores on the two parts may be regarded as an estimate of the reliability of the test half as long as the original test. From this, the reliability of the entire test can be estimated by means of the Spearman-Brown prophecy formula. Since division of a test into two halves may be made in many different ways, the reliability coefficient obtained by this method is not a constant value. It would be undesirable to split the test in the middle, because

the two parts might differ considerably in difficulty. The first half would be more novel to the subjects, the second half more influenced by fatigue factors and practice effects. The second half would be more difficult if the test item were arranged in ascending order of difficulty. The split-half method does not provide a satisfactory result on speed tests where many subjects fail to complete the entire test.

VALIDITY

The validity of a test is defined as "that aspect of a test which insures that it will actually measure what it claims to measure."[7] Validity is the most important function of a good test, for unless a test is valid it serves no useful purpose. A test must be valid for the purpose for which it it to be used. No single characteristic of a test can be labeled its "validity" in a general sense. Rather its validity must be considered with reference to the particular use to which it is to be put. A test satisfactorily valid for one purpose may be and often is worthless for another. To understand the full significance of a test score, then, it is necessary to know what performance the score predicts and how well it predicts that performance. This can be answered adequately only after a tryout of the test and a statistical analysis of the results.

A distinction is frequently made between "content" validity and "statistical" validity. A test is said to possess high content validity when each item in it has been drawn from undeniably important material from the subject field. A test is said to possess high statistical validity when it differentiates to a high degree between those who know more and those who know less. In other words, a valid test will arrange the students in order of their knowledge of the subject.

The content validity of a test is determined by the selection of questions, but the statistical validity can be discovered only from the data obtained by trying out the test.

Long and Sandiford[8] have the following suggestions for safeguarding the validity of a test.

1. The examination should be designed to measure a specific ability.

2. The examination should be based upon the curriculum which the teachers have been following. The curriculum is considered to be very fair common ground, for each teacher is supposed to have covered the course therein specified.

2. The examinations should be based upon the textbooks the pupils are supposed to have studied.

4. The examination should be a fair sampling of the subject. The fairest example is one that covers the entire course. Objective tests are able to provide a more adequate sampling than that obtained by essay-type tests.

5. The examination should be limited to the more important aspects of the subject. This selection is not an easy one and some of the ways in which it is attempted are:

(a) a consensus of opinion from experienced teachers;

(b) a study of old examination papers on the subject will provide a list of important topics;

(c) topics should fit the main objectives of the course;

(d) important topics are those which have the greatest residual social values. Topics which prove their worth in the after life of the individual are more important than those which are never used.

6. The examinations questions should be worded in clear, simple English. Ambiguity of statement and faulty sentence structure should be avoided.

7. The examination should be designed to encourage the pupil to do his best. The highest validity is secured by a steady performance on the part of the pupil. The steadiest performance is probably obtained when the pupil is working at his maximum capacity. Arrangement of the questions in order of increasing difficulty is perhaps the best method of securing the pupil's interest and effort.

The usual method of obtaining statistical validity of a test is by comparison of its results with those of an acknowledged criterion. The comparison is made by finding the coefficient of correlation between the two series of marks obtained when the test results and the criterion results are secured from the same group of subjects. Commonly used criteria against which to rate the test are school marks in the subject, average school mark, teachers' estimate of pupil ability, and another test. In Chapter V the procedure for determining the validity of the "Picture Test in Science" will make use of teachers' estimates of pupil ability correlated against the actual scores obtained on the test.

ITEM VALIDITY

Another aspect of testing with which this thesis is concerned is the validity of the test items. The validity of an item is defined as its effectiveness in discriminating between pupils at different levels of ability. Discriminative power of an item refers to the degree to which success or failure on the item is in itself indicative of ability in the general function presumed to be measured by the test as a whole.

It is true that the validity of the total test is a function of the validity of individual items. In many respects, however, it is desirable to give first attention to the validity of the individual test items than to the test as a whole. This is especially true when the test is an experimental edition as is the "Picture

Test in Science". This test contains eighty-one questions and in its present form it is of sufficient length to allow for questions that will be eliminated when the data are subjected to item analysis. Validation techniques will be used in this thesis to demonstrate how a long experimental test may be cut down to a shorter, more valid test.

Item-Validity Techniques. Many indices have been devised to analyze individual test items. Some of these reflect considerable statistical ingenuity and make possible appreciable savings of time if the number of items to be handled is large. Twenty-one methods of test item validation are described by Long and Sandiford[9].

(a) Bi-serial r. As this item-validity method will be dealt with completely in a later chapter only a formula summary will be given at the present time.

$$r_{bis} = \frac{M_2 - M_1}{\sigma} \times \frac{p\,q}{z}$$

in which:

M_2 = mean criterion score of group solving the item correctly.

M_1 = mean criterion score of group failing to solve the item correctly.

σ = standard deviation of all criterion scores.

p = proportion of total group solving the item correctly.

q = 1 - p = proportion of total group failing to solve the item correctly.

z = ordinate of the normal curve cutting off 'p' proportion of cases.

(b) The Vincent Method. The validity value is the percentage of those failing the item who have higher criterion scores than the median criterion score of those passing the item. For example, consider a test administered to 100 students with 30 failing a particular item. If, of these 30, 6 have

criterion scores higher than the median criterion score of the 70 who got the item correct, the Vincent validity value for the item will be 6/30 of 100 = 20. The smaller this margin of overlapping, the more sharply does the item discriminate between the good and poor students and so the higher is its validity. Usually the criterion score is the total score on the test. Vincent values range from 0 for the most valid items to 100 for the least valid items. Values greater than 50 indicate that the validity of the item is negative. Long and Sandiford claim that the Vincent method "enjoys the virtue of ease of computation but that it makes rather inadequate use of the data to which it is applied and sometimes leads to serious inexactitudes."[10]

(c) The Vincent Method Inverted. This method claims that just as good an indication of validity would be furnished by a technique which does just the reverse of the Vincent method, measuring the degree to which the 'passes' overlap the 'fails'. Thus, the Vincent inverted validity value for any item is the per cent of those passing the item who have criterion scores lower than the median criterion score of those who fail.

(d) The McCall Method. William A. McCall devised this validity technique for multiple-response items. The theory of this method is based on the assumption that the item which is most valid is one which best divides the candidates into widely separated groups. The ideal item would be one in which a certain response attracted only the best candidates, a second response attracted only the poorest candidates, and so on. The process of deriving the validity value of an item comprises two steps. The first consists in finding, for each of the possible responses, the average score of the candidates making that response. The second consists in calculating

the deviation of these averages from the average of the entire group. These deviations, considered without respect to sign, are weighted and averaged and the resulting mean deviation is the McCall measure of validity. The McCall formula is:

$$A.D.m_y = \frac{\Sigma f_x \left| Y^1_x \right|}{N}$$

in which:

f_x = frequency of a particular response group.

$\left| Y^1_x \right|$ = deviation without regard to sign of the mean criterion score of the particular response group from the mean criterion score of the entire group.

N = total number of cases.

The McCall method can be adapted to solve the more usual type of question in which items are scored dichotomously. This adaptation provides for the possibility of negative validity in an item, whereas no provision is made for this in the formula stated above.

(e) The McCall-Long-Bliss Method. This technique was developed by two of McCall's students in an effort to eliminate some of the inconsistencies to which the McCall method is liable. The McCall-Long-Bliss(M-L-B) formula is as follows:

$$V_{MLB} = \frac{(M_1 - M_2)\ f_1\ f_2 + (M_1 - M_3)\ f_1\ f_3 + \ldots\ldots\ldots}{\sigma\ N^2}$$

in which:

c $M_1, M_2, M_3, \ldots\ldots$ = the average criterion scores in order of size from highest to lowest of the various responses to this item.

$f_1, f_2, f_3 \ldots\ldots$ = the frequencies of the respective responses.

σ = the standard deviation of all criterion scores.

N = the number of criterion scores.

This measure, as compared with the McCall, is much more sensitive to minor changes in the grouping of criterion scores. The McCall method tends to favor items as they approach 50% difficulty but the M-L-B method tends to favor items which yield an even distribution of criterion scores among various responses.

(f) The Symonds' Index of Validity Method. By this method the same number of papers are selected from the top and the bottom of the class. Then the pupils' correct responses are tabulated on each item of the test. The measure of the goodness of the item is the difference between the number of correct responses for the good group and the poor group.

(g) The Symonds' Balance Method. Symonds investigated the relationship between validity and difficulty of an item. He showed that 'the items with which one can measure the ability of an individual most accurately are the items that he can do with 50% accuracy.'[11] A product is obtained for each item on the test by multiplying the number right by the number wrong. These products are then compared with the 50% level of difficulty to determine the most valid items.

(h) The Upper and Lower Halves Method. It is not known who first used the Upper and Lower Halves method. The procedure steps for this method are as follows:

1. arrange the papers in order of the size of the criterion scores beginning with the highest;
2. find the median;
3. using the median as the dividing line separate the papers into two groups;

4. examine each item of the test making the following three calculations:

a) the percentage of good pupils passing the item,

b) the percentage of poor pupils passing the item,

c) subtract b) from a).

According to this method the greater the difference the greater the validity of the item.

(i) The Upper and Lower Thirds Method. The Upper and Lower Thirds method is an improved modification of the Upper and Lower Halves method. The technique is exactly the same except that the item is evaluated by the difference between the number of passes for the highest and lowest one-third of the pupils.

(j) The Kelley Method. This method was used as a validation procedure on the 'Picture Test in Science' and a complete application of it will be found later in this discourse. The title is repeated here to complete the listing of methods outlined by Long and Sandiford.

(k) The Clark Method.[12] The Clark formula is:

$$V = \frac{P - D}{1 - D}$$

in which:

V = validity of the item.

D = proportion of the pupils who fail to answer the item correctly - the difficulty of the item.

P = Proportion of the criterion group who fail to answer the item correctly. The criterion group is the D proportion of the class having the lowest total scores.

For items of positive validity the range of V's is from 0 to 1. For items of negative validity the range of V's is from 0 to negative values whose limits

are conditioned by the difficulties of the items. It is interesting to note that the same validity will result if the formula is based on the number of passes.

(l) Cook's Index of Discrimination-A Method. This method consists in finding the percentage of pupils scoring an item wrongly in a group consisting of those who would mark it wrongly if the item had perfect discriminating power.

$$\text{Index A} = \frac{F}{E^1}$$ in which:

E = number of pupils in total group failing the item.

E^1 = number of lowest scoring pupils who would fail the item if it were perfect in discriminating power ($E^1 = E$)

F - number of pupils failing item in group E^1.

The higher the index, the greater the validity of the item. An index of 1.00 indicates perfect validity. The index remains positive even for negatively discriminating items. Index A has a strong tendency to give higher validity values to more difficult items.

(m) Cook's Index of Discrimination-B Method. This index of discrimination is identical with the Clark method. Long and Sandiford give proof of the identity of the two indices.

$$\text{Index B} = \frac{FH - E^2}{EH - E^2}$$

in which:

E = number of pupils in total group failing item.

E^1 = number of lowest scoring pupils who would fail the item if it were in perfect discriminating power ($E^1 = E$)

F = number of pupils failing item in group E^1.

H = total number of pupils.

(n) Cook's Index of Discrimination-C Method. This index is really a special form of the Upper and Lower technique.

$$\text{Index C} = \frac{N - M}{75}$$

in which:

M = per cent of accuracy for lowest one-fourth of group.

N = per cent of accuracy for highest one-fourth of group.

75 = number of percentiles of ability between the median score of these two groups.

(o) Cook's Index of Discrimination-D Method. This index consists in finding the ratio of the percent of accuracy with which the upper one-fourth of the pupils answer an item to the per cent of accuracy for the lowest one-fourth.

$$\text{Index D} = \frac{N}{M}$$

in which:

M = per cent of accuracy for the lowest one-fourth.

N = per cent of accuracy for the highest one-fourth.

(p) Lentz' Summation of the Agreement Method. According to Lentz "the item being evaluated is credited, in the case of each subject, with the number of responses to all items which agree with that subject's response to the item in question. The mean of the total agreeing scores constitutes the measure of the item's value."[13] The formula is:

$$C = \frac{N_{a_1} + N_{a_2} + N_{a_3} + \ldots\ldots\ldots N_{a_n}}{N}$$

in which:

C = the coefficient of validity.

Na_1 = total number of responses by the first subject which agree with that subject's response to the item being considered.

Na_2 = same applied to second subject.

N = total number of subjects responding to item.

(q) The Long Overlapping Method. This method is similar to the Vincent in its underlying principle but it makes more complete and more reliable use of the data presented. The Long method eliminates most of the inconsistencies of the Vincent method. The formula is:

$$V = 1 - \frac{2 \sum \text{"passes" below "fails"}}{(N_p)\ (N_f)}$$

in which:

N_p = the number of "passes"

N_f = the number of "fails".

(r) The Long Weighted Overlapping Method. This method has the advantage that items are selected through a technique which yields validity values which are uncorrelated with item difficulty. The item's chance of achieving high validity is not conditioned by its position in the difficulty scale. The formula is:

$$V = \frac{4(N_p)\ (N_f) \ - \ 8 \sum \text{"passes" below "fails"}}{N^2}$$

in which:

N = the total number of testees, other symbols same as in preceding method.

(s) The Correlation Ratio or Eta Method. The formula is:

$$N_{xy} = \frac{\sigma\ m_y}{\sigma_y}$$

in which:

m_y = mean criterion score of the group of pupils making a particular response.

σm_y = the standard deviation of these means (weighted) from the mean criterion score of the total group.

σy = the standard deviation of the criterion scores of the total group.

(t) Henry's Consistency of Performance Method. The procedure steps for this method are:

1. List the pupils in order of scores and divide the group into tenths.
2. Count the passes in each tenth.
3. Subtract the last tenth from the first to find the total decreases in the number of passes.
4. Proceeding from the first tenth to the last, find
   a) the number of intervals showing a decrease in the number of passes.
   b) the number of intervals showing no decrease.
   c) subtract b) from a).
5. Multiply the result obtained in step 3 by the result obtained in step 4. The higher the score, the greater the validity.

u. Henry's Mean Criterion Difference Method. This method consists in the evaluation of items on the basis of the difference between the average scores of the pupils having the item right and the average score of the pupils having the item wrong. Hence,

Mean Criterion Difference = Mean Right Score - Mean Wrong Score.

The greater the Mean Criterion Difference the more valid is the item.

Although only two of the above item-validity techniques are used in the following chapters, this summary of methods is provided to make this study a dictionary of item-validity procedures. Four additional methods, not included in Long and Sandiford, will be outlined and applied in Chapters VI and VII.

CHAPTER III

DESIGN OF THE STUDY

The Edmonton "Picture Test in Science" was used in this study to secure a body of data which would permit the application of the more promising statistical devices. The data gathered from the test will be used to show how test reliability, test validity, and item validity are determined.

This test was prepared by a committee of Edmonton elementary teachers under the chairmanship of Mr. T. Blacklock. An experimental edition of the test, consisting of eighty-one questions, was administered to a random sample of four hundred pupils in the Edmonton Public School system in May, 1950. The random sample of 400 was distributed over Grades IV to VII, 100 pupils in each grade. Four classrooms were selected at random for each grade. All the pupils in each classroom took the test but to secure the necessary 400 papers only 25 pupils were selected at random from each class. The teachers in charge of these selected classrooms marked the papers and these scores furnished the data for this study.

The test provided an excellent basis for statistical study. In the first place, the test is an experimental edition and as such it is much longer than is absolutely necessary. As a result, item-analysis is desirable in order to delete invalid items and reduce the test to a more manageable form. In the second place, it is necessary to discover the reliability and validity of the test to determine whether it would be advisable to publish the test in a final revised form.

Following is a detailed outline of the plan of this study, listing the statistical procedures which will be used in succeeding chapters. The

purpose is to illustrate the employment of the statistical techniques and to make a comparison of the results obtained.

Reliability. The reliability of the 'Picture Test in Science' will be determined by the following methods:

1. the Split-Half method corrected by the Spearman Brown prophecy formula,
2. the Rational Equivalence method using the Froelich adaptation of the Kuder-Richardson formula.
3. the Analysis of Variance method as outlined by Hoyt.

Validity. The validity of the 'Picture Test in Science' will be determined by correlating the test scores and teachers' estimates of science ability for 100 Grade VII pupils.

Item Validity. Item Validity for the 81 items on the 'Picture Test in Science' will be determined by the following methods:

1. the Davis Item-Analysis technique which provides both difficulty and discrimination indices,
2. the Flanagan Co-efficient,
3. the Kelly Item Validity method,
4. the Standard Error of the Difference between two percentages,
5. the Lawshe Nomograph,
6. the Bi-serial r method on items 21, 35, 43, 52 and 70.

The following tables show tabulation of data which were compiled for the purpose of this study:

1. Table I, Calculation of Mean and Standard Deviation for 400 scores, and a summary of means and standard deviations for each grade;
2. Table II, Tabulation of Choices on Each Item for Item Validity calculation. The top and bottom 27% of the test papers were used for this tabulation. Correct choices are underlined and changed to percentages.

Table I is included in this chapter to present a comprehensive picture of the 400 test scores. The mean and standard deviation is shown for the entire sample and also for each of the separate grades. Certain item analysis methods will make use of the data in this table.

Table II is the main tabulation table of this study and most of the item analysis methods which follow will make extensive use of this information. It supplies a complete tabulation of choices on all items for the 27% of 400 pupils scoring highest and lowest on the 'Picture Test in Science'.

TABLE I. CALCULATION OF MEAN AND STANDARD DEVIATION FOR 400 SCORES (GRADES IV - VII)

Scores	f	x'	fx'	fx'^2
70 - 74	1	+5	5	25
65 - 69	5	+4	20	80
60 - 64	26	+3	78	234
55 - 59	45	+2	90	180
50 - 54	77	1	77	77
45 - 49	114	0	+270	
40 - 44	76	-1	-76	76
35 - 39	33	-2	-66	132
30 - 34	12	-3	-36	108
25 - 29	11	-4	-44	176
			-222	
	N = 400		270-222 = 48	1088

i = 5

A.M. = 47.00

c = 48 ÷ 400 = .12

c^2 = .0144

ci = .12 x 5 = .6

M = 47.00 + .6 = 47.6

$$S.D. = \sqrt{\frac{\Sigma fx'^2}{N} - c^2} \times i$$

$$= \sqrt{\frac{1088}{400} - .0144} \times 5$$

$$= \sqrt{2.72 - .0144} \times 5$$

$$= \sqrt{2.7056} \times 5$$

$$= 1.64 \times 5$$

$$= 8.20$$

	GR. IV	GR. V	GR. VI	GR. VII
M	40.70	46.25	50.15	53.3
σ	6.65	6.55	6.45	7.25

TABLE II. TABULATION OF CHOICES ON EACH ITEM FOR ITEM VALIDITY COMPUTATION BASED ON 27% SCORING HIGHEST AND LOWEST. (27% OF 400 PUPILS = 108)

Ques. No.	27% Scoring Highest							27% Scoring Lowest						
	Tally of Choices					Tot.	%	Tally of Choices					Tot.	%
	1	2	3	4	0			1	2	3	4	0		
1	1	2	105	0	0	108	97	4	3	99	2	0	108	92
2	0	99	9	0	0	108	92	2	74	28	2	2	108	69
3	1	4	103	0	0	108	95	12	5	83	7	1	108	77
4	80	1	1	26	0	108	74	46	9	6	45	2	108	43
5	1	10	7	90	0	108	83	8	13	31	52	4	108	48
6	1	93	10	4	0	108	86	10	55	34	7	2	108	51
7	107	0	0	1	0	108	99	103	0	1	4	0	108	95
8	27	14	54	13	0	108	50	11	49	36	11	1	108	32
9	8	46	9	45	0	108	43	17	18	12	58	3	108	17
10	10	1	3	94	0	108	87	17	1	5	82	3	108	76
11	5	5	43	54	1	108	40	15	10	20	59	4	108	19
12	57	7	3	40	1	108	37	47	14	2	39	6	108	36
13	8	70	14	15	1	108	65	12	42	20	29	5	108	39
14	10	28	61	9	0	108	56	14	34	43	17	0	108	40
15	26	36	8	33	5	108	31	13	66	16	12	1	108	11
16	100	2	5	1	0	108	93	77	4	19	6	2	108	71
17	58	1	44	5	0	108	54	41	1	56	8	2	108	38
18	2	28	0	78	0	108	72	5	28	3	72	0	108	67
19	1	81	10	15	1	108	75	3	61	18	26	0	108	56
20	0	3	102	3	0	108	94	2	4	79	23	0	108	73

TABLE II - (Continued)

Ques. No.	1	2	3	4	0	Tot.	%	1	2	3	4	0	Tot.	%
21	60	36	5	6	1	108	32	60	26	8	13	1	108	24
22	2	93	5	7	1	108	86	6	70	7	25	0	108	65
23	0	0	101	6	1	108	94	2	1	92	13	0	108	85
24	1	37	36	31	3	108	34	16	15	28	48	1	108	14
25	41	6	2	59	0	108	38	34	22	10	41	1	108	31
26	1	0	106	1	0	108	98	22	4	73	6	3	108	68
27	3	5	100	0	0	108	93	7	17	78	5	1	108	72
28	1	94	3	10	0	108	87	1	77	3	19	8	108	71
29	7	15	83	3	0	108	77	23	20	51	12	2	108	47
30	11	12	33	51	1	108	47	14	37	34	20	3	108	19
31	58	9	26	15	0	108	54	11	13	51	30	3	108	10
32	1	8	87	9	3	108	81	6	28	56	15	3	108	52
33	1	1	11	95	0	108	88	9	2	31	61	5	108	56
34	4	62	17	24	1	108	57	9	33	23	38	5	108	31
35	35	39	9	24	1	108	36	17	24	12	46	9	108	22
36	3	0	97	8	0	108	90	5	15	65	18	5	108	60
37	4	6	2	94	2	108	87	15	22	8	60	3	108	56
38	1	8	13	85	1	108	79	3	24	30	49	2	108	45
39	54	7	33	13	1	108	50	34	11	32	21	10	108	31
40	87	0	8	12	1	108	7	99	2	1	5	1	108	1
41	15	1	0	92	0	108	85	25	8	4	69	2	108	64
42	10	1	87	10	0	108	81	26	4	45	32	1	108	42

TABLE II - (Continued)

Ques. No.	1	2	3	4	0	Tot.	%	1	2	3	4	0	Tot.	%
43	69	38	0	1	0	108	35	77	22	2	5	2	108	20
44	0	5	102	1	0	108	94	0	29	73	5	1	108	68
45	0	80	25	3	0	108	74	5	51	47	5	0	108	47
46	9	25	5	69	0	108	64	8	35	10	55	0	108	51
47	8	18	61	21	0	108	56	21	42	20	23	2	108	19
48	0	107	1	0	0	108	99	2	98	4	3	1	108	91
49	19	8	16	65	0	108	60	33	13	24	33	5	108	31
50	34	21	18	31	4	108	31	14	28	33	25	8	108	13
51	73	2	6	26	1	108	68	53	7	15	27	6	108	49
52	46	44	5	13	0	108	41	37	24	8	35	4	108	22
53	50	33	8	16	1	108	46	26	32	12	36	2	108	24
54	52	30	12	14	0	108	28	34	38	11	22	3	108	35
55	16	10	26	52	4	108	9	10	8	30	56	4	108	7
56	17	22	38	31	0	108	29	19	35	33	18	3	108	17
57	24	21	1	61	1	108	56	36	17	7	41	7	108	38
58	64	38	0	4	2	108	59	55	34	6	9	4	108	51
59	1	101	0	6	0	108	94	2	75	10	20	1	108	69
60	27	69	7	5	0	108	64	49	45	10	4	0	108	42
61	4	7	1	96	0	108	89	16	8	5	78	1	108	72
62	102	3	2	0	1	108	94	95	3	2	6	2	108	88
63	21	2	80	5	0	108	74	32	5	64	6	1	108	59
64	1	0	0	107	0	108	99	10	0	0	97	1	108	90

TABLE II - (Continued)

Ques. No.	1	2	3	4	0	Tot.	%	1	2	3	4	0	Tot.	%
65	16	0	91	1	0	108	84	32	5	65	3	3	108	60
66	5	2	21	80	0	108	74	6	7	35	57	3	108	53
67	4	3	76	22	3	108	70	19	10	26	45	8	108	24
68	2	23	5	77	1	108	71	12	28	7	59	2	108	55
69	44	35	22	5	2	108	32	61	10	18	12	7	108	9
70	7	51	35	13	2	108	32	22	41	25	15	5	108	23
71	17	6	14	70	1	108	65	42	16	9	28	13	108	26
72	47	40	4	16	1	108	44	31	29	5	30	13	108	29
73	89	2	3	13	1	108	82	65	5	4	28	6	108	60
74	1	103	4	0	0	108	95	1	85	16	2	4	108	79
75	4	13	20	71	0	108	66	5	28	23	48	4	108	44
76	5	0	0	103	0	108	95	13	1	3	88	3	108	81
77	8	100	0	0	0	108	93	24	67	5	10	2	108	62
78	1	2	8	96	1	108	89	12	9	13	69	5	108	64
79	0	106	1	1	0	108	98	2	95	3	6	2	108	88
80	89	1	7	10	1	108	82	66	2	21	17	2	108	61
81	28	52	9	17	2	108	48	50	35	4	16	3	108	32

CHAPTER IV

THE RELIABILITY OF THE 'PICTURE TEST IN SCIENCE'

As outlined in Chapter III, the reliability of the 'Picture Test in Science' will be determined by three separate and distinct methods. These are the Split-Half Method corrected by the Spearman-Brown prophecy formula, the Method of Rational Equivalence using the Froelich adaption of the Richardson-Kuder formula, and the Analysis of Variance Method as outlined by Hoyt.

The Split-Half Method. The split-half division which was used on the 'Picture Test in Science' was the scores on the odd-numbered and even-numbered items. The distribution of the 400 test scored for this even-odd distribution is shown in Table III. The frequency distribution of these scores is shown on the scatter diagram in Figure 1. The computational work shown on Figure 1 provides a reliability coefficient of .64 for the test. This correlation between the scores on the two parts may be regarded as an estimate of the reliability of a test half as long as the original test. From this estimate of the reliability of half of the test, the reliability of the entire test can be estimated. This is done by means of the Spearman-Brown prophecy formula, which takes into account the reduction in length of the test due to splitting the number of items in half and estimates what the reliability would be if each half were doubled in length. The formula is:

$$r_t = \frac{2r_{12}}{1 + r_{12}}$$

where r_t = the reliability coefficient of the total test.

r_{12} = the correlation between the scores on the halves of the test.

Application of the above formula to the reliability coefficient of .64 secured

from Figure 1 will prophesy the reliability coefficient of the entire 'Picture Test in Science' to be .78 .

The Method of Rational Equivalence. Kuder and Richardson[14] developed an index which approximates the coefficient of reliability as determined by the split-halves technique when used in conjunction with the Spearman-Brown formula, but which does not involve the computation of a Pearsonian "r".

The main advantages of this method are that it requires no odd-even rescoring, no second administration of test, and no duplicate form to be constructed and administered. The main disadvantage is that the computational work requires a great deal of time. The index was designed for use with an electric test scoring machine.

The "Rational Equivalence" formula is:

$$r = \frac{n}{(n-1)} \times \frac{\sigma_t^2 - \sum pq}{\sigma_t^2}$$

in which:

r = reliability coefficient of whole test.

n = number of items in the test.

σ_t = the S.D. of the test scores.

p = the proportion of the group answering a test item correctly.

q = (1-p) = the proportion of the group answering a test item incorrectly.

pq requires the computation of p and q for each item in the test, multiplying p and q for each item, followed by the summation for all items.

The method of rational equivalence stresses the intercorrelations of the items in the test and the correlations of the items with the test as a whole.

A simple adaption to the above "rational equivalence" formula has been devised by Froelich.[15]

$$r = \frac{n\ \sigma_t^2 - M\ (n-M)}{\sigma_t^2\ (n-1)}$$

in which:

r = reliability of the whole test.

n = number of items in the test .

σ_t = S.D. of the test scores.

M = the mean of the test scores.

This formula is a labor saver and supplies a satisfactory approximation to the test's reliability even when the test items cover a wide range of ability. This formula will underestimate that obtained by the split-half technique, but it can be considered as providing a minimum of reliability. On the 'Picture Test in Science' this formula yielded a reliability coefficient of .72 compared with .78 secured by the Split-Half method.

Froelich claims that his formula "will yield sufficiently accurate results to warrant its use by classroom teachers."[16] Teachers can use this formula to determine quickly the reliability of short objective classroom examinations because it requires a minimum of time and statistical background.

The Analysis of Variance Method. A very significant contribution to procedure in estimating reliability was made by Hoyt[17] who developed a formula for estimating the reliability of a test by analysis of variance methods. Tables IV and V show the tabulations which are necessary for the analysis of variance process. The formula for estimating the reliability of a test by analysis of variance procedure is:

$$r = \frac{a - c}{a}$$

in which:

r = reliability coefficient of test.

a = mean of squares between individuals.

c = mean of squares of residual.

This formula provided a reliability coefficient of .78 on the 'Picture Test in Science' which is exactly the same as that secured by the split-half method.

Procedure for Analysis of Variance.

1. Arrange the test papers in descending order. The 400 papers for the Picture test in Science were arranged in order of raw scores with the high scores of 68 at top and the low score of 24 on the bottom.

2. Construct a student-item chart (Table IV). The numbers across the top relate to the 81 individual test items on the Picture Test in Science. The numbers in the left-hand column relate to the test scores, arranged in order of mark, of the 400 pupils who took the test. The items which each of the 400 pupils marked correctly were checked with an "X"; e.g., student 1 had the high score of 68, and an "X" was placed under each of the 68 out of 81 items that he had correct on his test paper. The same procedure was followed for each paper down to paper 400 with the low score of 24. Because of the large number of test papers it was necessary to record the results on 20 separate tabulation sheets measuring 8 1/2" x 24". The totals from each of these pages were transferred to a summary page of the same size. Table IV is only able to show a small part of this summary.

3. Add the individual columns and rows. These are labeled t_1 and p_1 in Table IV. Identical results should be found for each addition. For the

Picture Test in Science the total was 18725.

4. Square each t_i and p_i value. These results are posted in column t_i^2 and row p_i^2. Add the column and the row. These totals must be added very carefully as there is no identity check as in step 3.

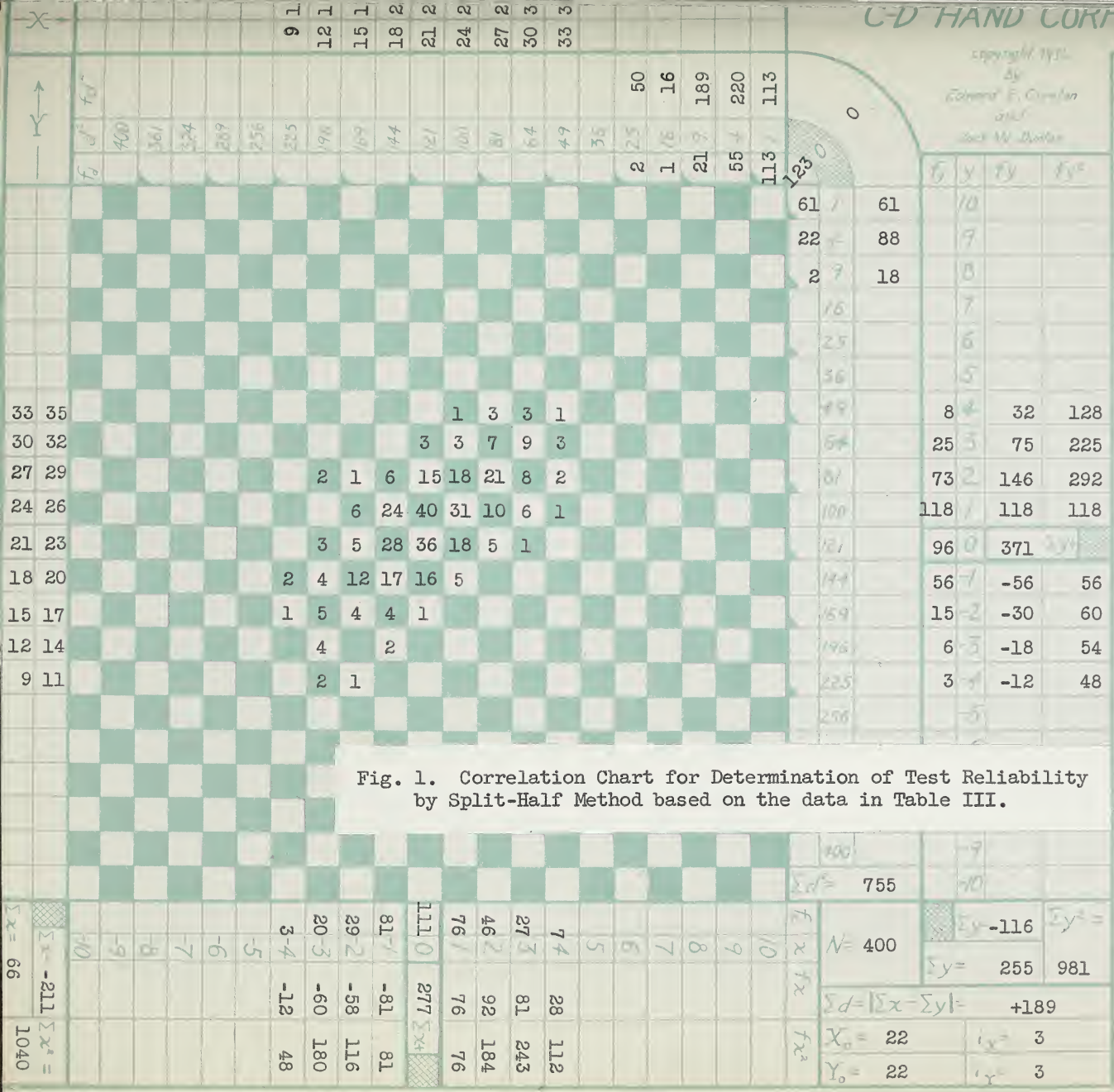
5. Use the data in Table IV to construct Analysis of Variance Table V.

TABLE III. DISTRIBUTION OF 400 TEST SCORES INTO EVEN-ODD CLASSIFICATION FOR DETERMINATION OF RELIABILITY COEFFICIENT BY SPLIT-HALF TECHNIQUE

No.	EVEN	ODD	No.	EVEN	ODD	No.	EVEN	ODD	No.	EVEN	ODD	No.	EVEN	ODD
1	35	33	41	29	27	81	23	30	121	23	27	161	22	26
2	30	34	42	29	28	82	26	27	122	28	22	162	24	24
3	32	33	43	29	27	83	30	23	123	26	23	163	23	25
4	31	33	44	30	27	84	28	25	124	27	22	164	24	24
5	32	32	45	28	29	85	25	28	125	28	22	165	24	23
6	29	35	46	31	25	86	28	25	126	23	27	166	24	23
7	33	31	47	26	30	87	26	26	127	24	26	167	25	22
8	33	30	48	28	29	88	28	24	128	22	27	168	22	25
9	33	31	49	28	28	89	29	24	129	26	24	169	25	22
10	32	30	50	27	30	90	26	27	130	26	24	170	23	24
11	30	31	51	27	29	91	27	25	131	24	25	171	28	19
12	34	28	52	27	29	92	27	25	132	27	23	172	27	21
13	29	33	53	27	28	93	26	26	133	24	26	173	26	21
14	31	31	54	28	28	94	27	25	134	24	26	174	25	22
15	31	31	55	26	30	95	25	26	135	26	24	175	26	22
16	33	28	56	30	25	96	30	22	136	26	22	176	21	26
17	30	31	57	28	27	97	28	24	137	24	25	177	26	21
18	32	29	58	29	27	98	28	23	138	24	25	178	24	24
19	32	29	59	29	27	99	30	22	139	28	20	179	25	23
20	30	31	60	27	29	100	25	26	140	25	24	180	22	24
21	29	31	61	26	30	101	26	25	141	27	21	181	25	21
22	31	30	62	29	25	102	24	27	142	26	23	182	27	20
23	31	28	63	28	26	103	27	25	143	28	21	183	21	26
24	28	32	64	28	27	104	27	24	144	26	22	184	27	19
25	33	27	65	25	30	105	25	25	145	24	24	185	25	21
26	29	30	66	25	29	106	27	24	146	24	25	186	26	20
27	30	30	67	28	27	107	27	23	147	22	27	187	21	25
28	29	30	68	28	26	108	25	25	148	23	26	188	24	23
29	28	31	69	27	27	109	25	25	149	26	22	189	24	23
30	31	27	70	26	28	110	29	22	150	25	23	190	21	25
31	28	30	71	28	27	111	25	26	151	22	27	191	24	22
32	26	32	72	27	27	112	28	23	152	25	24	192	26	20
33	27	32	73	28	26	113	24	27	153	23	26	193	24	23
34	25	34	74	26	29	114	28	23	154	25	23	194	21	25
35	31	28	75	29	26	115	25	25	155	24	25	195	25	22
36	31	27	76	26	29	116	26	25	156	24	24	196	24	22
37	28	29	77	26	27	117	27	23	157	25	23	197	23	24
38	25	32	78	29	24	118	27	23	158	26	23	198	25	21
39	34	24	79	27	27	119	25	25	159	23	26	199	22	24
40	32	25	80	28	25	120	25	25	160	27	21	200	26	20

TABLE III - (Continued)

No.	EVEN	ODD	No.	EVEN	ODD	No.	EVEN	ODD	No.	EVEN	ODD	No.	EVEN	ODD
201	25	22	241	26	19	281	20	23	321	22	18	361	22	14
202	24	22	242	23	22	282	21	22	322	27	13	362	17	18
203	26	20	243	25	20	283	24	19	323	21	19	363	20	16
204	27	19	244	24	20	284	22	21	324	22	19	364	21	14
205	23	23	245	26	19	285	23	19	325	21	20	365	15	20
206	22	24	246	24	20	286	20	23	326	19	21	366	20	15
207	23	23	247	24	20	287	23	19	327	22	18	367	17	17
208	22	23	248	23	21	288	26	16	328	20	19	368	21	14
209	24	22	249	21	24	289	21	22	329	21	18	369	19	16
210	25	21	250	23	21	290	22	20	330	20	19	370	19	16
211	22	24	251	22	23	291	25	18	331	22	17	371	14	20
212	23	22	252	25	19	292	18	23	332	19	20	372	20	15
213	25	21	253	21	23	293	26	16	333	21	19	373	19	16
214	22	23	254	20	24	294	24	18	334	22	18	374	18	17
215	26	19	255	24	19	295	26	16	335	22	17	375	19	15
216	26	20	256	22	22	296	23	19	336	19	19	376	17	17
217	25	21	257	22	22	297	22	20	337	19	19	377	15	19
218	26	20	258	23	20	298	21	20	338	19	20	378	20	14
219	24	21	259	24	20	299	20	22	339	21	18	379	19	14
220	25	21	260	23	21	300	25	17	340	20	18	380	19	15
221	25	21	261	27	18	301	21	21	341	21	18	381	17	17
222	24	22	262	25	19	302	19	22	342	21	17	382	19	13
223	24	22	263	21	22	303	21	20	343	19	18	383	14	18
224	22	23	264	20	24	304	21	21	344	19	19	384	20	12
225	20	26	265	22	21	305	19	22	345	20	18	385	19	11
226	23	22	266	18	25	306	21	19	346	22	16	386	17	14
227	23	21	267	27	16	307	18	23	347	19	18	387	16	14
228	23	22	268	23	20	308	22	19	348	17	20	388	16	14
229	25	20	269	21	23	309	24	16	349	17	21	389	15	15
230	22	23	270	23	21	310	22	19	350	18	19	390	16	13
231	23	21	271	23	21	311	21	20	351	18	18	391	18	9
232	24	21	272	23	19	312	21	20	352	21	16	392	14	14
233	23	21	273	20	22	313	20	21	353	19	17	393	15	13
234	23	21	274	19	23	314	27	14	354	21	16	394	13	14
235	22	22	275	23	19	315	20	21	355	18	19	395	11	15
236	22	23	276	22	21	316	18	22	356	20	17	396	16	10
237	25	20	277	24	19	317	23	18	357	18	19	397	11	14
238	24	21	278	20	23	318	20	21	358	19	18	398	13	12
239	22	23	279	24	19	319	20	21	359	18	19	399	12	13
240	20	24	280	22	21	320	24	16	360	20	17	400	11	14


C-D HAND CORRELATION CHART
Study Picture Test in Science
Subject Random Sample - 400
Var X Odd Scores
Var Y Even Scores
Investigator D. Cooney 26/2/51
Computer D. Cooney 26/2/51
COMPUTATIONS
N = 400
Σy = 255
Σy² = 981
Σx = -116
Σd = |Σx - Σy| = +189
X₀ = 22
i_x = 3
Y₀ = 22
i_y = 3
Fig. 1. Correlation Chart for Determination of Test Reliability by Split-Half Method based on the data in Table III.

C-D HAND CORRELATION CHART

Copyright 1952 by
Edward E. Cureton and
Jack W. Dunlap

Distributed By
The Psychological Corporation
522 Fifth Ave., New York City

The C-D Hand Correlator Chart is designed to facilitate the computation of the Pearson product-moment correlation coefficient and certain related functions without the use of a calculating machine. It is arranged to insure the maximum accuracy with the minimum of computational work.

DIRECTIONS

Fill in the data in the upper right corner.

For each variable, find the lowest score and the highest score; subtract the lowest from the highest. This difference is the range. Divide the range by 15 (disregard any fractional remainder) and take the quotient or the next lower convenient whole number as the grouping interval. Convenient grouping intervals are 1, 2, 3, 5, 7, 10, 15, 20, etc. It is well for convenience to make the lower limit of the lowest group a multiple of the grouping interval. Record X_0 and i_x, the grouping intervals selected, in the spaces provided in the lower right corner of the chart.

In the two columns to the extreme left of the chart, fill in the boundaries of the X intervals. Variable Y should be the variable for which the range of the two is to be so considered. It is best, though not necessary, to arrange the distribution so that the center of gravity is slightly above it. If it is usually the case that there are more than 21 intervals, in a similar manner fill in the boundaries of the X intervals in the two rows at the top of the chart, arranging the distribution for convenience near the center of a field to the right. The smaller values of the variables should always come at the bottom and left respectively. Record X_0 and Y_0, the midpoints of the zero intervals, in the spaces provided at the bottom of the chart. The zero interval for the Y's lies in the center row of the scarlet diagonals, i.e., the row which leads to the right through 12, 0, $2y_1$, and a dark square. The zero interval for the X's lies in the column directly above 0, $2x_1$, and another dark square at the bottom of the chart.

Plot the paired scores on the scatter diagram.

Count the number of scores in each row and record each sum in the cell that lies in the column headed f_y (cell 3y). And the entries in this column and record the sum, N, in the space provided at the bottom of the chart. Turn the chart so that the left side becomes the bottom, and again count the frequencies in the rows (formerly columns), recording the sum in each case in the appropriate cell of the column headed f_x. Add the entries in this column. Their sum should equal N (check).

Turn the chart right side up again. Count the number of scores in each of the diagonals running from lower left to upper right. The checkerboard design facilitates the addition, as each diagonal will be either white or blue. Each blue diagonal leads to a space (at top of right) having a blue triangle in its lower left corner. Record the sums of scores in the diagonals in the spaces immediately above and to the right of the scarlet diagonals or checkerboard, in the round-the-corner column headed f_d. Entries along the top should be written with the chart turned so that the right side becomes the bottom. Add the entries in the round-the-corner column. Their sum should again be equal to N.

Multiply each score in the f_x column by the primed value next to it in the y column, and record the products in the fy column. There will of course be no value in the corner of the row of 0's and the values below the center will all be negative. Next multiply each value in the fy column by the corresponding value in the y column, and record the products in the fy^2 column. There will again be no entry in the center, but all the entries will be positive. Add the upper half of the fy column and record the sum in the box at the center labelled $\Sigma y+$. Add the lower half of this column and record the sum in the box labelled $\Sigma y-$. Add algebraically the values $\Sigma y+$ and $\Sigma y-$ and record this sum as Σy. Now add the fy^2 column and record the sum in the box labeled Σy^2.

Turn the chart so that the left side becomes the bottom and tabulate Σx and Σx^2 in a similar manner.

If Σx and Σy both have the same sign, subtract the one which is numerically smaller from the one which is numerically larger, and record the difference as Σd in the space provided at the bottom of the chart, giving it the sign of the larger. If Σx and Σy are of opposite sign, add their numerical values, and record with positive sign as Σd. Now multiply each value of f_d in the round-the-corner column by the corresponding value of d' printed next to it, and record in the adjacent column headed fd'. Add all the values in the round-the-corner fd' column, and record the sum in the box labeled $\Sigma d'$.

This completes the preliminary computations. The final work is indicated in the column at the right of the chart headed COMPUTATIONS. Each necessary step is indicated in the column headed Operation. The result of each step should be recorded in the column headed Value. The column headed Result gives the new name of each entry in the Value column. Each value required at any stage is indicated in the Operation column, may be found among the preliminary computations or among the previous entries in the Result column.

The work below the results is optional. Operations are indicated for obtaining the standard error of the correlation coefficient, the sigma and standard deviations of the original series in their original score units σ_x, the product-moment coefficient, and the two regression coefficients.

All additions and subtractions indicated in the COMPUTATIONS columns are to be taken as algebraic additions or subtractions. If the correlation is negative, $(N\sigma_x)^2$ will have a value greater than the sum of $(N\sigma_y)^2$ and $(N\sigma_d)^2$, and the numerator of the correlation coefficient will be negative. In this case the product-moment coefficient and the two regression coefficients will be negative also.

TABLE IV. SUMMARY FOR STUDENT-ITEM CHART TO PROVIDE DATA FOR ANALYSIS OF VARIANCE

	1	2	3	4	5	6	7	8	9	10	11	12	13	72	73	74	75	76	77	78	79	80	81		t_i		t_i^2
1	x	x	x	x	x	x	x		x	x	x	x	x	x	x	x		x	x	x	x	x			68		4624
2	x	x	x	x	x	x	x		x	x	x	x	x		x	x	x	x	x	x	x	x	x		65		4225
3	x	x	x	x	x	x	x	x	x	x	x				x	x	x	x	x	x	x	x			65		4225
4	x	x	x	x	x	x	x		x	x	x		x	x	x	x		x	x	x	x	x	x		65		4225
5	x	x	x	x	x	x	x	x				x	x	x		x	x	x	x	x	x	x	x		65		4225
6	x	x	x	x	x	x	x		x	x	x		x		x		x	x	x	x	x	x			64		4096
7	x	x	x	x	x		x	x	x	x	x		x	x	x	x		x	x	x	x	x			64		4096
8	x	x	x	x	x	x	x	x		x	x		x		x	x	x	x	x	x	x	x	x		64		4096
395	x		x	x	x	x	x			x					x						x	x			26		676
396				x	x		x	x		x			x				x				x	x			26		676
397	x		x				x						x			x	x			x	x				25		625
398	x	x	x	x			x	x		x												x			25		625
399	x		x				x									x		x		x	x				25		625
400	x						x					x		x		x				x	x				24		576
																								ΣP_i	Σt_i		Σt_i^2
Pi	373	327	340	333	273	283	383	162	111	333	109	149	211	147	237	252	220	365	311	310	372	274	173	18725		903339
ΣP_i^2	139129	106929	115600	54289	74529	80089	146689	26244	12321	110889	11881	22201	44521	21609	88209	123904	48400	133225	96721	96100	138384	75076	29929	ΣP_i^2		5062241	

TABLE V. ANALYSIS OF VARIANCE TABLE FOR PICTURE TEST IN SCIENCE WITH COMPUTED VALUES

SOURCE OF VARIANCE	DEGREES OF FREEDOM (df)	SUM OF SQUARES	MEAN OF SQUARES	F	HYPOTHESIS TESTED
BETWEEN INDIVIDUALS	K = No. of Individuals K - 1 = 400 - 1 = 399	$\frac{1}{n}\sum t_i^2 - \frac{(\sum t_i)^2}{nk}$ $= \frac{1}{81} \times 903339 - \frac{(18725)^2}{81 \times 400}$ $= \frac{903339}{81} - \frac{350625625}{32400}$ = 331	$\frac{\text{SUM OF SQUARES}}{df}$ $= \frac{331}{399} = .83$ (a)	$\frac{a}{c}$ $= \frac{.83}{.18}$ = 4.6	Table for "F" (Table 4, p.65) Lindquist Statistical Analysis in Educational Research. Entering the Table with 80 and 399 Degrees of Freedom, the "F" value is found to be 1.54 at the 1% level. Both "F" values of this Table (4.6 and 127.2) are greater than 1.54, and hence we can assume that the test is significant.
BETWEEN ITEMS	N = No. of Items N - 1 = 81 - 1 = 80	$\frac{1}{k}\sum p_i^2 - \frac{(\sum t_i)^2}{nk}$ $= \frac{1}{400} \times 5062241 - \frac{(18725)^2}{81 \times 400}$ $= \frac{5062241}{400} - \frac{350625625}{32400}$ = 12656 - 10822 = 1834	$\frac{\text{SUM OF SQUARES}}{df}$ $= \frac{1834}{80} = 22.9$ (b)	$\frac{b}{c}$ $= \frac{22.9}{.18}$ = 127.2	
RESIDUAL	(N-1)(K-1) = 80 x 399 = 31920	TOTAL -(Between individuals + between items) = 7903 - (331 + 1834) = 7903 - 2165 = 5738	$\frac{\text{SUM OF SQUARES}}{df}$ $= \frac{5738}{31920} = .18$ (c)		COMPUTATION FOR RELIABILITY COEFFICIENT OF THE TEST $r_{tt} = \frac{a - c}{a}$ $= \frac{.83 - .18}{.83}$ $= \frac{.78}{}$
TOTAL	NK - 1 = (81x400)-1 = 32400 - 1 = 32399	$\text{TOTAL} = \frac{(\sum t_i)(nk - \sum t_i)}{nk}$ $= \frac{18725(400 \times 81 - 18725)}{400 \times 81}$ = 7903			

CHAPTER V

THE VALIDITY OF THE PICTURE TEST IN SCIENCE

Statistical validity of the test refers to the mathematical processes for determining the degree to which the test agrees or correlates with some criterion which is set up as an acceptable measure of the ability in question. The criterion used in this study for correlation against actual test scores to determine test validity was the teachers' estimates of pupil ability in science. As this information was incomplete on all the 400 pupils of the random sample test, validity procedures had to be limited to the 100 Grade VII pupils. It is likely that the criterion possesses weaknesses, but it is only being used to illustrate how test validity may be obtained.

Table VI provides a summary of the 100 Grade VII test scores with their corresponding criterion scores which were used for the determination of validity for the Picture Test in Science. The data of Table VI is used in Figure 2 where a scattergram is shown of the two sets of scores and the computational procedures required for securing a correlation coefficient for test validity.

Figure 2 shows the coefficient of test validity to be .27. This is a low validity coefficient but we cannot expect to find a very high one because the coefficients of correlation between the test scores and the criterion are limited by the unreliability of both the test and the criterion.

Because the test possesses low validity as measured by a comparison of test scores against teachers' estimates, it is important that an analysis be made to determine specific weaknesses in the test. Item analysis procedures based on this test will be outlined in the next two chapters.

TABLE VI. 100 GRADE VII TEST SCORES WITH CORRESPONDING CRITERION SCORES FOR DETERMINATION OF TEST VALIDITY

	X	Y		X	Y		X	Y		X	Y
Pupil No.	Crit. Score	Test Score	Pupil No.	Crit. Score	Test Score	Pupil No.	Crit. Score	Test Score	Pupil No.	Crit. Score	Test Score
1	50	52	26	75	57	51	75	62	76	45	59
2	85	60	27	65	65	52	60	54	77	65	60
3	40	53	28	70	55	53	60	49	78	55	62
4	75	64	29	60	65	54	85	64	79	55	61
5	80	56	30	40	56	55	65	51	80	75	52
6	80	68	31	45	65	56	60	55	81	55	49
7	80	55	32	38	46	57	59	55	82	51	59
8	60	56	33	56	58	58	57	37	83	69	58
9	44	56	34	30	46	59	63	62	84	62	49
10	59	53	35	42	53	60	33	52	85	44	50
11	37	53	36	60	56	61	29	49	86	66	57
12	62	48	37	80	59	62	29	43	87	50	65
13	52	35	38	44	53	63	55	43	88	70	64
14	65	54	39	70	64	64	65	55	89	85	63
15	60	49	40	65	60	65	60	44	90	85	60
16	60	60	41	65	45	66	85	51	91	60	52
17	60	47	42	50	41	67	50	48	92	55	46
18	60	36	43	70	40	68	85	59	93	55	57
19	50	47	44	50	46	69	65	45	94	72	50
20	65	47	45	41	43	70	72	50	95	50	47
21	56	45	46	57	46	71	59	46	96	74	57
22	64	44	47	51	47	72	84	57	97	88	55
23	59	44	48	50	45	73	55	49	98	52	45
24	48	50	49	67	46	74	73	56	99	54	41
25	76	56	50	69	51	75	59	48	100	55	46

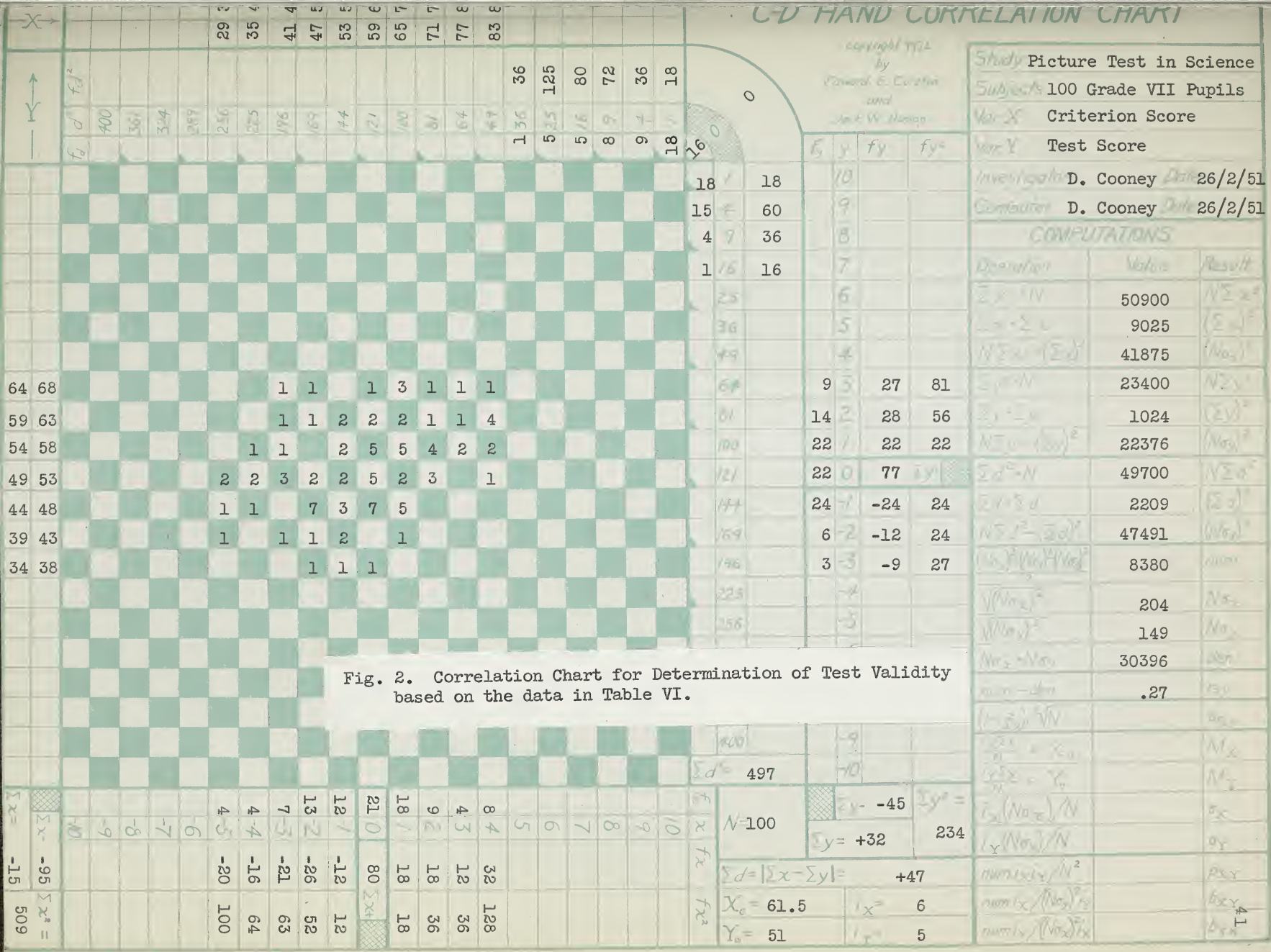


Fig. 2. Correlation Chart for Determination of Test Validity based on the data in Table VI.

C-D HAND CORRELATION CHART

Copyright 1952-63

Edward E. Cureton and
Jack W. Dunlap
Consultants To
The Psychological Corporation
522 Fifth Ave., New York City

The C-D Hand Correlation Chart is designed to facilitate the computation of the Pearson product-moment correlation coefficient and certain related functions without the aid of a calculating machine. It is arranged to assure the maximum accuracy with the minimum of computational work.

DIRECTIONS

Fill in the data in the upper right corner.

For each variable, find the lowest score and the highest score. Subtract the lowest from the highest. This difference is the range. Divide the range by 10, dropping any fractional remainder, and take the quotient as the next lower convenient whole number as the grouping interval. Convenient grouping intervals are 1, 2, 3, 5, 7, 10, 15, 20, etc. It is well for convenience to make the lower limit of the lowest group a multiple of the grouping interval. Record i_x and i_y, the grouping intervals selected, in the spaces provided at the bottom of the chart.

In the two columns to the extreme left of the chart fill in the boundaries of the X intervals. Variable X should be the one having fewer, or rather not the one having more, intervals. It is best, though not necessary, to arrange the distribution for convenience so that the scores are in the center of the chart. If, as is usually the case, there are less than 11 intervals, fill in a similar manner, but in the boundaries of the X-intervals in the two rows at the top of the chart, arranging the distribution for convenience so that the scores are in the center of the chart. The smaller values of the variables should always come at the bottom and left, respectively. Record X_0 and Y_0, the midpoints of the zero intervals, in the spaces provided at the bottom of the chart. The zero interval for the X's lies in the center row of the scatter diagram, i.e., the row which lends itself through 121, O, Σy_{+1}, and a dark square. The zero interval for the Y's lies in the column directly above O, Σx_{+1}, and another dark square at the bottom of the chart.

Take the paired scores on the scatter diagram.

Count the number of cases in each row and record their sum in the cell that lies in the column headed f_y, (not fy). Add the entries in this column and record the sum, N, in the space provided at the bottom of the chart. Turn the chart so that the left side becomes the bottom, and recount the frequencies in the rows. Formerly column f_y recording the sum in each case in the appropriate cell of the column headed f_x. Add the entries in this column. They sum should equal N (check).

Turn the chart right side up again. Count the number of scores in each of the diagonals running from lower left to upper right. The checkerboard design facilitates the addition, as each diagonal will be either white or blue. Each cell adjacent (not to a space (at top or right) having a blue triangle in its lower left corner. Record the sums of scores in the diagonals in the spaces immediately above and to the right of the corner triangles or checkerboard, in the round-the-corner column headed f_u. Entries along the top should be written with the chart turned so that the right side becomes the bottom. Add the entries in the round-the-corner column. Their sum should again be equal to N.

Multiply each entry in the f_u column by the printed value next to it in the u column, and record the products in the fu column. There will of course be no value in the center where $u = 0$, and the values below the center will all be negative. Next multiply each value in the fu column by the corresponding value in the v column, and record the products in the fu' column. There will again be no entry in the center, but all the entries will be positive. Add the upper half of the fv column and record the sum in the box at the center labelled $\Sigma uv+$. Add the lower half of this column and record the sum in the box labelled $\Sigma uv-$. Add *algebraically* the values $\Sigma uv+$ and $\Sigma uv-$, and record this sum as Σuv. Now add the fu' column and record the sum in the box labelled $\Sigma u'$.

Turn the chart so that the left side becomes the bottom and calculate Σx and $\Sigma x'$ in a similar manner.

If Σx and Σu have the same sign, subtract the one which is numerically smaller from the one which is numerically larger and record the difference as Σd in the space provided at the bottom of the chart, always with positive sign. If Σx and Σu are of opposite sign, add their numerical values, and record with positive sign as Σd. Next multiply each value of f_x in the first round-the-corner column by the corresponding value of d' printed next to it, and record in the adjacent column headed fd'. Add all the values in the round-the-corner fd' column, and record the sum at the bottom as $\Sigma d'$.

This completes the preliminary computations. The final work is indicated in the columns at the right of the chart headed COMPUTATIONS. Each necessary step is indicated in the column headed Operation. The result of each step should be recorded in the column headed Value. The column headed Result gives the page where each entry in the Value column. Each value required at any stage, as indicated in the Operation column, may be found among the preliminary computations or among the previous entries in the Result column.

The work below the result r_{xy} is optional. Operations are indicated for obtaining the standard error of the correlation coefficient, the means and standard deviations of the original series in their original scale units, $b_{y.x}$, the product-moment coefficient of regression, and the two regression coefficients.

All additions and subtractions indicated in the COMPUTATIONS column are to be taken as algebraic additions or subtractions. If the correlation is negative, $(\Sigma u')^2$ will have a value greater than the sum of $(\Sigma u_+)^2$ and $(\Sigma u_-)^2$, and the numerator of the correlation coefficient will be negative. In this case the product-moment coefficient and the two regression coefficients will be negative also.

CHAPTER VI

THE DAVIS ITEM-ANALYSIS TECHNIQUE

Many techniques of item analysis have been proposed in recent years, but "perhaps the most satisfactory index of discrimination for achievement test purposes is one devised by Davis."[18] The Davis method is outlined in a booklet entitled "Item Analysis Data - Their Computation, Interpretation, and Use in Test Construction", by Frederick B. Davis, Professor of Psychology at the George Peabody College for Teachers.

The procedure outlined by Davis is designed to overcome the disadvantage of other techniques. An Item-Analysis Chart accompanies the booklet and it is possible to secure both difficulty and discrimination indices in one operation. Computational procedures which may be used are outlined fully and formulas are presented in an easy-to-handle form. Davis claims that his method has many advantages over other procedures.

1. It is designed for maximum convenience in test construction, whereas nomographs and tables in use at present are not.

2. Difficulty indices are supplied, thus overcoming a serious deficiency of present tables.

3. Discrimination indices are supplied in a more usable form than those secured from other tables.

4. It is based on sound statistical procedures and its use demands a minimum of labor.

The Davis Item-Analysis Chart has another important advantage - one that merits a separate paragraph for emphasis. Many item-analysis procedures employ the relationship between each individual item and some criterion

other than the total score. Such procedures are very valuable but frequently it is a very difficult matter to obtain a suitable criterion. Davis claims that "the new item-analysis chart in this bulletin is equally convenient for use whether the criterion variable be the total score on the test in which the item is included or some other continuous quantitative series."[19] Furthermore, the use of the chart does not depend on the employment of the recommended formulas for computing proportions of successes. Usable difficulty and discrimination indices can be obtained from the chart if it is entered with proportions computed in any one of several ways.

The item-analysis chart is composed of cells each containing two numbers arranged vertically. The top value in each cell in the chart is the discrimination index. This provides a numerical indication of the amount of discrimination each individual item possesses with respect to a designated criterion variable which is usually the total score on the test in which the item is included. The bottom value in each cell in the chart is the difficulty index which provides a numerical indication of the difficulty level of each individual item relative to the other items administered to the same sample of subjects.

The Difficulty Index. "The difficulty of any item may be defined as the proportion of a certain sample marking the item correctly, or it may be defined as the proportion of a certain sample who actually know the answer to an item."[20] As the Edmonton 'Picture Test in Science' is the multi-choice test, the possibility always exists that a certain proportion of subjects may mark the correct answer by chance. Each of the eighty-one questions on the test is a four-choice item. If the subjects do not know the correct answer to a

question and if they all guess, it is most likely that one-fourth of them will mark the correct answer. A correction for chance is employed in the Davis formula by subtracting from the number of subjects who mark the item correctly the number of subjects who marked the item incorrectly divided by one less than the number of the choices in the item. Ordinarily, this procedure tends to overcorrect for chance because some subjects choose incorrect answers on the basis of the misinformation rather than on the basis of sheer guessing. Davis claims that this tendency to overcorrect for chance may, in some measure, be offset by the fact that the incorrect items are rarely equally attractive to the subject. Horst[21] has shown that the proportion of subjects selecting the most popular incorrect answer is equal to the proportion marking the correct answer by chance alone.

The values of the difficulty indices in the chart were obtained by using the Kelley-Wood Table of the Normal Probability Integral.[22] The range of the indices is from 1 to 99.

The Discrimination Index. Perhaps the most satisfactory measure of relationship for internal consistency item-analysis is the bi-serial correlation coefficient between each item and the total score on the test. Such data would probably be used more often if their computations were less laborious. A short-cut procedure has been devised to approximate them with a minimum of labor. Kelley[23] demonstrated that the lowest and the highest 27% of the sample are the optimum groups for use when item-analysis data are being prepared. Flanagan[24], using Kelley's procedure, constructed a table of the values of the product-moment coefficient of correlation corresponding to given proportions of successes. Davis made use of these values in his item

analysis chart. The values in his table were converted into discrimination indices by means of Table X, which includes equivalent values of product-moment r, Fisher's z, and Davis' discrimination index. According to the Davis chart, items with discrimination indices above 20 will ordinarily possess sufficient discrimination power for use in most achievement and aptitude tests.[25]

Davis Computation Procedures. The figures required to enter the item-analysis chart consist of the proportions of success in the highest and lowest 27% of the sample. After a set of test papers has been scored the papers are arranged in order of rank with respect to the score that is to be used as the criterion for item analysis. The highest and lowest 27% of the sample are then removed. If more than one school grade is represented in the sample, as was the case in the 'Picture Test in Science', the highest and lowest 27% of each grade should be selected separately to ensure that pupils will be selected from each grade. The highest 27% for each grade should then be combined to form the high-scoring group for the item-analysis and similarly for the low-scoring group. Tabulation should then be made in both groups for the number of pupils that marked the correct answer to the item, the number that marked any wrong answer, and the number that left the question unanswered. The percentages required to enter the item-analysis chart are now computed from this data and difficulty and discrimination indices are secured from the one operation.

If an item is so difficult that fewer than would be expected by chance in both high-scoring and low-scoring groups answer it correctly, both figures needed to enter the chart will be negative. To compensate for this, serviceable indices are obtained if the chart is entered as though the proportions were

positive and then the signs of the indices found were reversed.

If for a given item the signs of the two proportions used to enter the chart are different there is no straightforward way of using the item-analysis chart. Davis makes use of Table VII which provides a corrected percentage for entering the chart when the percentage of successes in the lowest 27% of the sample is 0 or negative. If the percentage of successes in the highest 27% is above 99, the percentage used to enter the item-analysis chart must first be altered by reference to Table VIII. Table IX provides a summary of difficulty and descrimination indices for all items on the 'Picture Test in Science' as determined by the Davis method.

The writer of this study disagrees with the Davis method when use of Table VII is made to secure corrected percentages when the percentages of successes in the lowest 27% of the sample is 0 or negative. On five items in the 'Picture Test in Science' where it was necessary to use this procedure very acceptable discrimination indices were secured from the item-analysis chart. These results did not agree with other item-analysis procedures outlined in the next chapter. Finally, bi-serial r calculations were carried out on these five doubtful items and in each case the discrimination index was low. Table XVI shows the bi-serial r calculations for item 21 which provides a validity index of .06 as compared with a Davis index of 26.5 . A complete summary of these findings is contained in Table XVII.

TABLE VII. PERCENTAGES SUGGESTED FOR ENTERING THE DAVIS ITEM-ANALYSIS CHART WHEN THE PERCENTAGE OF SUCCESSES IN THE LOWEST 27% OF THE SAMPLE IS ZERO OR NEGATIVE

Value of $N_L - NR_L$ in formula (13)	Percentage Suggested for Entering Chart
34 or more	1
20 - 33	2
15 - 19	3
12 - 14	4
10 - 11	5
8 - 9	6
7	7
6	8
5	10

Frederick B. Davis, Item - Analysis Data, p.35.

TABLE VIII. PERCENTAGES SUGGESTED FOR ENTERING THE DAVIS ITEM-ANALYSIS CHART WHEN THE PERCENTAGE OF SUCCESSES IN THE HIGHEST 27% OF THE SAMPLE IS CLOSEST TO 100

Value of $N_H - NR_H$ in formula (12)	Percentage Suggested for Entering Chart
34 or more	99
20 - 23	98
15 - 19	97
12 - 14	96
10 - 11	95
8 - 9	94
7	93
6	92
5	90

Frederick B. Davis, Item - Analysis Data, p.36.

TABLE IX. SUMMARY OF DISCRIMINATION AND DIFFICULTY INDICES OBTAINED FROM DAVIS ITEM-ANALYSIS CHART

Item No.	Discrimination Index	Difficulty Index	Item No.	Discrimination Index	Difficulty Index
1	12.5	80.5	21	26.5	17
2	24.5	63	22	20.5	59.5
3	25	69	23	15	74
4	27	48	24	29	19
5	29	53	25	10.5	26.5
6	30.5	55.5	26	41	66
7	19	88	27	23	66
8	19	34	28	10.5	64.5
9	38	25.5	29	24	50
10	10	65	30	44	29
11	36.5	24.5	31	49	32
12	-1	30.5	32	26	54
13	22	43.5	33	27	58
14	16.5	39.5	34	28	37
15	25	16	35	31	20.5
16	23.5	65.5	36	27	60
17	15	38	37	29.5	58
18	4.5	55	38	27.5	51
19	16	52.5	39	18	34.5
20	27	66.5	40	6.5	-37.5

TABLE IX - (Concluded)

41	19	59.5	61	17	63.5
42	35	49	62	9.5	76.5
43	30	20	63	12	53.5
44	31	64.5	64	28	82
45	23	49	65	21	58
46	11	46	66	17	52
47	51	33	67	62.5	40
48	28	82	68	13	51
49	30	38	69	25	16
50	25	16	70	25	16
51	13	48	71	41	39
52	36.5	24.5	72	15	30.5
53	44	29	73	18	58
54	-17.5	22.5	74	21	72
55	3	-33.5	75	17.5	45.5
56	16.5	10.5	76	19	73
57	15.5	40	77	31.5	61
58	7	46	78	23	61
59	28	65	79	19	79
60	20	43	80	19	58
			81	18.5	33.5

TABLE X. EQUIVALENT VALUES OF PRODUCT-MOMENT "r" (FLANAGAN), FISHER'S "Z", AND DAVIS DISCRIMINATION INDICES

DAVIS	Z	r	DAVIS	Z	r	DAVIS	Z	r
0	.00	.00	34	.56	.51	67	1.11	.80
1	.02	.02	35	.58	.52	68	1.13	.81
2	.04	.04	36	.60	.54	69	1.14	.81
3	.05	.05	37	.61	.55	70	1.16	.82
4	.07	.07	38	.63	.56	71	1.18	.83
5	.08	.08	39	.65	.57	72	1.19	.83
6	.10	.10	40	.66	.58	73	1.21	.84
7	.12	.12	41	.68	.59	74	1.23	.84
8	.14	.14	42	.69	.60	75	1.24	.85
9	.15	.15	43	.71	.61	76	1.26	.85
10	.17	.17	44	.73	.62	77	1.28	.86
11	.18	.18	45	.74	.63	78	1.29	.86
12	.20	.20	46	.76	.64	79	1.31	.86
13	.22	.22	47	.78	.65	80	1.33	.87
14	.23	.23	48	.79	.66	81	1.34	.87
15	.25	.24	49	.81	.67	82	1.36	.88
16	.27	.26	50	.83	.68	83	1.38	.88
17	.28	.27	51	.85	.69	84	1.39	.88
18	.30	.29	52	.87	.70	85	1.41	.89
19	.32	.31	53	.88	.70	86	1.43	.89
20	.33	.32	54	.89	.71	87	1.44	.89
21	.35	.34	55	.91	.72	88	1.46	.90
22	.36	.35	56	.93	.73	89	1.48	.90
23	.38	.36	57	.95	.74	90	1.49	.90
24	.40	.38	58	.97	.75	91	1.51	.91
25	.41	.39	59	.98	.75	92	1.53	.91
26	.43	.40	60	1.00	.76	93	1.54	.91
27	.44	.41	61	1.02	.77	94	1.56	.92
28	.46	.43	62	1.03	.77	95	1.58	.92
29	.48	.45	63	1.05	.78	96	1.59	.92
30	.50	.46	64	1.07	.79	97	1.61	.92
31	.51	.47	65	1.08	.79	98	1.63	.93
32	.53	.48	66	1.10	.80	99	1.64	.93
33	.55	.50				100	1.66	.93

CHAPTER VII

OTHER ITEM VALIDITY TECHNIQUES

As stated in the last chapter the 1950 edition of the Encyclopedia of Educational Research is used as the authority for stating that the Davis method provides the most satisfactory index of discrimination. In order to compare the Davis with other methods five additional procedures for item-analysis will be outlined in this chapter. The first four methods described will use the highest and lowest 27% of the class for comparison groups. This is the same basic data which was used for the Davis method. Each of these four methods will supply item-analysis results on all 81 questions of the "Picture Test in Science". The fifth method outlined in this chapter will be the bi-serial r. As laborious calculations are required for this method it will only be used, as intimated in the preceding chapter, on the five items which appear to have doubtful ratings by the Davis method.

The Flanagan Item Analysis Method. Flanagan[26] has prepared a chart which can be used to determine the validity of test items. The chart is based on Tables VIII and IX in Tables for Statisticians and Biometricians, Part II, edited by Karl Pearson. The chart shows the values of the product-moment coefficient of correlation corresponding to given proportions of successes in the upper and lower 27% of the criterion groups. This method was devised to reduce the large amount of time and effort required by bi-serial r calculation. Flanagan claims that the results obtained from the use of his chart have been found to be satisfactory approximations to the bi-serial coefficients. The

coefficients in Flanagan's table range from -.93 to .93: higher values of "r" do not appear because percentages of the high-scoring and low-scoring groups smaller than 1% and larger than 99% were not considered.

The only available chart in the University of Alberta Library for obtaining Flanagan coefficients is on page 678 of Volume XXX of the Journal of Educational Psychology. This chart is small and difficult to use and it is practically impossible to secure indices on extremely high or low proportions. Table XI provides a summary of item validities for the "Picture Test in Science" secured by use of the Flanagan chart. The Flanagan method checks item for item with the Davis method with the exception of the five items on the Davis method which appear to have doubtful classification.

The Kelley Item Validity Method. The validity of an item by this method is taken as the distance, in sigma units, between the ordinates which cut off the percentage proportions of the upper and lower 27% which answer the item correctly from the area of the normal probability curve. The work of computation for any particular item involves the following steps:

1. find what proportion of the upper group answers the item correctly;
2. find what proportion of the lower group answers the item correctly;
3. from appropriate tables of the normal probability curve, find the position in sigma values of the ordinate cutting off the proportion of cases found in step I;
4. find the position, in sigma values of the ordinate cutting off the proportion of cases found in step 2;

5. subtract the result yielded in step 4 from the result yielded in step 3 and the remainder will represent the validity value of the item as derived by the Kelley method.

In the actual operation of this method a difficulty arises through the fact that the normal curve is asymptotic at each end to the base line and so extends indefinitely in both directions. It would follow from this that if an item is passed by 100 per cent of the lower group its validity value would be infinity, a distinction which it obviously does not deserve. Long and Sandiford suggest that "the curve be truncated at either end in such a manner that the highest 1 per cent and the lowest 1 per cent of the area be assumed to occupy a distance along the base line equal to that occupied by the second highest 1 per cent and the second lowest 1 per cent."[27] They also suggest that the confusion of positive and negative deviations can be eliminated by taking the lower end of the scale as the point of reference. Table XIII is built on the basis of the above suggestions and the use of such a table greatly expedites the determination of item validities.

This technique purports to assign validity values which have no tendency to be influenced by the factor of item-difficulty. Consider two items on Table XII, items No. 1 and 43. No. 1 was passed by 97% of the upper group and 92% of the lower group, and No. 43 was passed by 35% of the upper group and 20% of the lower group. The difference between the percentages is 5 for question 1 and 15 for question 43, hence by the Upper vs Lower method question 43 would be 3 times as valid as question 1. But by the Kelley method the validity of item 1 is 4.47

- 4.00 or .47, while that for item 43 is 2.20 - 1.75 or .45 which is approximately the same. The assumption underlying this is that a difference of 5 between the percentages for one item is as difficult to achieve as a difference of 15 between the percentages of the other item.

The Kelley technique is employed for item analysis by the Co-operative Test Service of the American Council of Education. Dr. Wood, the Director of the Co-operative Test Service makes use of a graphic device to simplify the labour of computation and set out the validities in such a manner that their comparative values can readily be seen by inspection. This device is illustrated in Figure 3. The percentages are converted into quarter-sigma units and laid off on the horizontal axis for the lower group and on the vertical axis for the upper group. Items are then plotted in the positions to which their percentages assign them. The validity of an item is measured by the distance of its position above the diagonal line. Item 26 is the most valid, whereas item 54 has negative validity.

This chart discriminates neither as minutely nor as accurately as does Table XIII, but this weakness is not serious because the test-maker is interested in validation only in so far as it will enable him to select the best and discard the poorest from a large number of trial items.

The Standard Error of the Difference Between Two Percentages

The percentages obtained on each item for the upper and lower 27% for use with Davis Item Analysis may further be utilized by determining whether the difference is significant. This is done by finding the percentage difference between the two groups and

dividing by the standard error of the difference. It is necessary to do this computation for each of the items. The resulting critical ratio obtained by this division can then be checked against a probability table to determine its extent of discrimination.

For example, the percentage difference between the two groups on Item 13 is 65-39 = 26. The standard error of the difference is 6.5. Dividing the percentage difference between the two groups by the standard error of the difference the critical ratio for this item is $26 \div 6.5 = 4.0$.

The critical ratio is then compared with the "t" value obtained from a probability table [28] to determine whether the result is significant or not. Using 107 (N-1) degrees of freedom to enter the table we find that an item must have a critical ratio of 1.98 to be significant at the 5% level. Therefore our example item with a CR of 4.0 exceeds the required ratio of 1.98 and we conclude that the ratio is significant. The item is said to discriminate and its degree of discrimination can be determined by comparing it with the critical ratios obtained for other items on the test.

Table XIV presents a complete summary of data for each item on the "Picture Test in Science". The last column shows the Critical Ratios (C.R.'s), and using the 5% level of significance the following questions would have to be eliminated from the test as they fail to discriminate: 7, 12, 18, 21, 25, 46, 54, 55, 58, 62 and 70. By using the more exacting and highly desirable 1% level the Critical Ratio required is 2.63, and the following additional items would fail to possess the necessary discrimination values: 1, 10, 14, 17, 23, 35, 40, 43, 56, 63, 68, 72, and 81.

The Lawshe Item-Analysis Nomograph. Lawshe[29] has developed a nomograph (Figure 4) that can be used for estimating the validity of individual test items. It simplifies the procedure described in the previous section in that it is not necessary to divide the percentage difference by the standard error of the difference. The nomograph is designed to do this calculation.

First, divide the test papers into high and low 27% groups. Second, compute the percentages of each group marking the item correctly. Next, locate these percentage points on the left and right sides of the nomograph. Connect these two points with a ruler and the discrimination value for the item will be the point on the middle line where it is crossed by the ruler. Negative signs always appear when there is a reversal, that is, when the poor group does better than the good group.

Table XV shows the frequency distribution of discrimination values for the items on the "Picture Test in Science". Lawshe suggests that items having a D-value less than 0.4 should be eliminated or revised. "Experience has indicated that in the average situation D-values less than 0.3 or 0.4 seldom contribute to a more reliable test.[30] Using 0.4 as the dividing line would eliminate 15 items. Table XVIII shows that if 0.5 were used as the dividing line the Lawshe method would compare very favorably with the other methods used in this study. This method shows discrepancies in certain instances but the use of the nomograph will save considerable time and for most purposes the results will be sufficiently precise.

The Bi-serial r Method. This is a method of correlation applicable to data in which one variable is quantitative and continuous, with the other variable presented in a dichotomous classification. For

validation of test items, the pupils can be measured in the first variable, but can only be classified into two categories in the second variable. The correlation between a set of scores and a two-category classification cannot be found by the ordinary product-moment formula or by the rank-difference method. The bi-serial method of correlation is dependent on two fundamental assumptions. The first of these is that the variable for which we have a dichotomous division is actually distributed normally. The second assumption is that the relationship between the two variables is linear. Each item on a test will yield its own particular bi-serial r, which can be considered a measure of that item's validity. The higher the coefficient, the better does the item predict ability in the criterion trait. Bi-serial r varies between the limits of + 1 and - 1. Its formula, with definitions of the symbols are as follows:

$$r_{bis} = \frac{M_2 - M_1}{\sigma} \times \frac{pq}{z}$$

M_2 = mean criterion score of group solving the item correctly.

M_1 = mean criterion score of groups failing to solve the item correctly.

σ = standard deviation of all criterion scores.

p = proportion of total group solving the item correctly.

$q = 1 - p$ = proportion of total group failing to solve the item correctly.

z = ordinate of the normal curve cutting off "p" proportion of cases.

The calculation of bi-serial r for item No. 21 from the "Picture Test in Science" is illustrated in Table XVI.

TABLE XI. SUMMARY OF DISCRIMINATION COEFFICIENTS OBTAINED BY FLANAGAN METHOD

Item No.	Corr. Coeff.	Item No.	Corr. Coeff.	Item No.	Corr. Coeff.	Item No.	Corr. Coeff.
1	.20	21	.09	41	.27	61	.25
2	.35	22	.28	42	.41	62	.14
3	.35	23	.20	43	.19	63	.17
4	.33	24	.27	44	.40	64	.50
5	.37	25	.08	45	.28	65	.29
6	.41	26	.55	46	.24	66	.23
7	.30	27	.35	47	.40	67	.46
8	.19	28	.23	48	.50	68	.17
9	.30	29	.32	49	.30	69	.35
10	.16	30	.32	50	.26	70	.12
11	.25	31	.51	51	.20	71	.40
12	.01	32	.32	52	.21	72	.16
13	.27	33	.38	53	.24	73	.26
14	.17	34	.27	54	-.08	74	.34
15	.30	35	.16	55	.05	75	.23
16	.35	36	.40	56	.16	76	.30
17	.17	37	.37	57	.19	77	.44
18	.06	38	.36	58	.08	78	.34
19	.20	39	.20	59	.40	79	.34
20	.36	40	---	60	.23	80	.25
						81	.18

TABLE XII. SUMMARY OF DATA FOR KELLY ITEM VALIDITY METHOD

Item	Percentage Upper Group	Percentage Lower Group	Percentages Converted to Sigma Distances - Area of Normal Curve		Difference
1	97	92	4.47	4.00	.47
2	92	69	4.00	3.09	.91
3	95	77	4.23	3.33	.90
4	74	43	3.23	2.41	.82
5	83	48	3.54	2.54	1.00
6	86	51	3.67	2.62	1.05
7	99	95	4.91	4.23	.68
8	50	32	2.59	2.12	.47
9	43	17	2.41	1.64	.77
10	87	76	3.72	3.30	.42
11	40	19	2.34	1.71	.63
12	37	36	2.26	2.23	.03
13	65	39	2.98	2.31	.67
14	56	40	2.74	2.34	.40
15	31	11	2.09	1.36	.73
16	93	71	4.07	3.14	.93
17	54	38	2.69	2.28	.41
18	72	67	3.17	3.03	.14
19	75	56	3.26	2.74	.52
20	94	73	4.14	3.20	.94
21	32	24	2.12	1.88	.24
22	86	65	3.67	2.98	.69
23	94	85	4.14	3.63	.51
24	34	14	2.18	1.51	.67
25	38	31	2.28	2.09	.19
26	98	68	4.64	3.06	1.58
27	93	72	4.07	3.17	.90
28	87	71	3.72	3.14	.58
29	77	47	3.33	2.51	.82
30	47	19	2.51	1.71	.80
31	54	10	2.69	1.31	1.38
32	81	52	3.47	2.64	.83
33	88	56	3.77	2.74	1.03
34	57	31	2.77	2.09	.68
35	36	22	2.23	1.82	.41
36	90	60	3.87	2.84	1.03
37	87	56	3.72	2.74	.98
38	79	45	3.40	2.46	.94
39	50	31	2.59	2.09	.50
40	7	1	1.11	.27	.84

TABLE XII - (Concluded)

41	85	64	3.63	2.95	.68
42	81	42	3.47	2.39	1.08
43	35	20	2.20	1.75	.45
44	94	68	4.14	3.06	1.08
45	74	47	3.23	2.51	.72
46	64	51	2.95	2.62	.33
47	56	19	2.74	1.71	1.03
48	99	91	4.91	3.93	.98
49	60	31	2.84	2.09	.75
50	31	13	2.09	1.46	.63
51	68	49	3.06	2.56	.50
52	41	22	2.36	1.82	.54
53	46	24	2.49	1.88	.61
54	28	35	2.01	2.20	-.19
55	9	7	1.25	1.11	.14
56	29	17	2.04	1.64	.40
57	56	38	2.74	2.28	.46
58	59	51	2.82	2.62	.20
59	94	69	4.14	3.09	1.05
60	64	42	2.95	2.39	.56
61	89	72	3.82	3.17	.65
62	94	88	4.14	3.77	.37
63	74	59	3.23	2.82	.41
64	99	90	4.91	3.87	1.04
65	84	60	3.58	2.84	.74
66	74	53	3.23	2.67	.56
67	70	24	3.11	1.88	1.23
68	71	55	3.14	2.72	.42
69	32	9	2.12	1.25	.87
70	32	23	2.12	1.85	.27
71	65	26	2.98	1.95	1.03
72	44	29	2.44	2.04	.40
73	82	60	3.51	2.84	.67
74	95	79	4.23	3.40	.83
75	66	44	3.00	2.44	.56
76	95	81	4.23	3.47	.76
77	93	62	4.07	2.90	1.17
78	89	64	3.82	2.95	.87
79	98	88	4.64	3.77	.87
80	82	61	3.51	2.87	.64
81	48	32	2.54	2.12	.42

TABLE XIII. PERCENTAGES OF AREA OF THE NORMAL AREAS, WITH SIGMA DISTANCES FROM THE LOWER END OF THE CURVE OF THE CORRESPONDING ORDINATES (CURVE TRUNCATED AT EITHER END AT A POINT 2.59 σ FROM THE MEAN)

100	5.18	66	3.03	32	2.12
99	4.91	65	2.98	31	2.09
98	4.64	64	2.95	30	2.07
97	4.47	63	2.92	29	2.04
96	4.37	62	2.90	28	2.01
95	4.23	61	2.87	27	1.98
94	4.14	60	2.84	26	1.95
93	4.07	59	2.82	25	1.92
92	4.00	58	2.79	24	1.88
91	3.93	57	2.77	23	1.85
90	3.87	56	2.74	22	1.82
89	3.82	55	2.72	21	1.78
88	3.77	54	2.69	20	1.75
87	3.72	53	2.67	19	1.71
86	3.67	52	2.64	18	1.69
85	3.63	51	2.62	17	1.64
84	3.58	50	2.59	16	1.60
83	3.54	49	2.56	15	1.55
82	3.51	48	2.54	14	1.51
81	3.47	47	2.51	13	1.46
80	3.43	46	2.49	12	1.41
79	3.40	45	2.46	11	1.36
78	3.36	44	2.44	10	1.31
77	3.33	43	2.41	9	1.25
76	3.30	42	2.39	8	1.18
75	3.26	41	2.36	7	1.11
74	3.23	40	2.34	6	1.04
73	3.20	39	2.31	5	.95
72	3.17	38	2.28	4	.84
71	3.14	37	2.26	3	.71
70	3.11	36	2.23	2	.54
69	3.09	35	2.20	1	.27
68	3.06	34	2.18	0	.00
67	3.03	33	2.15		

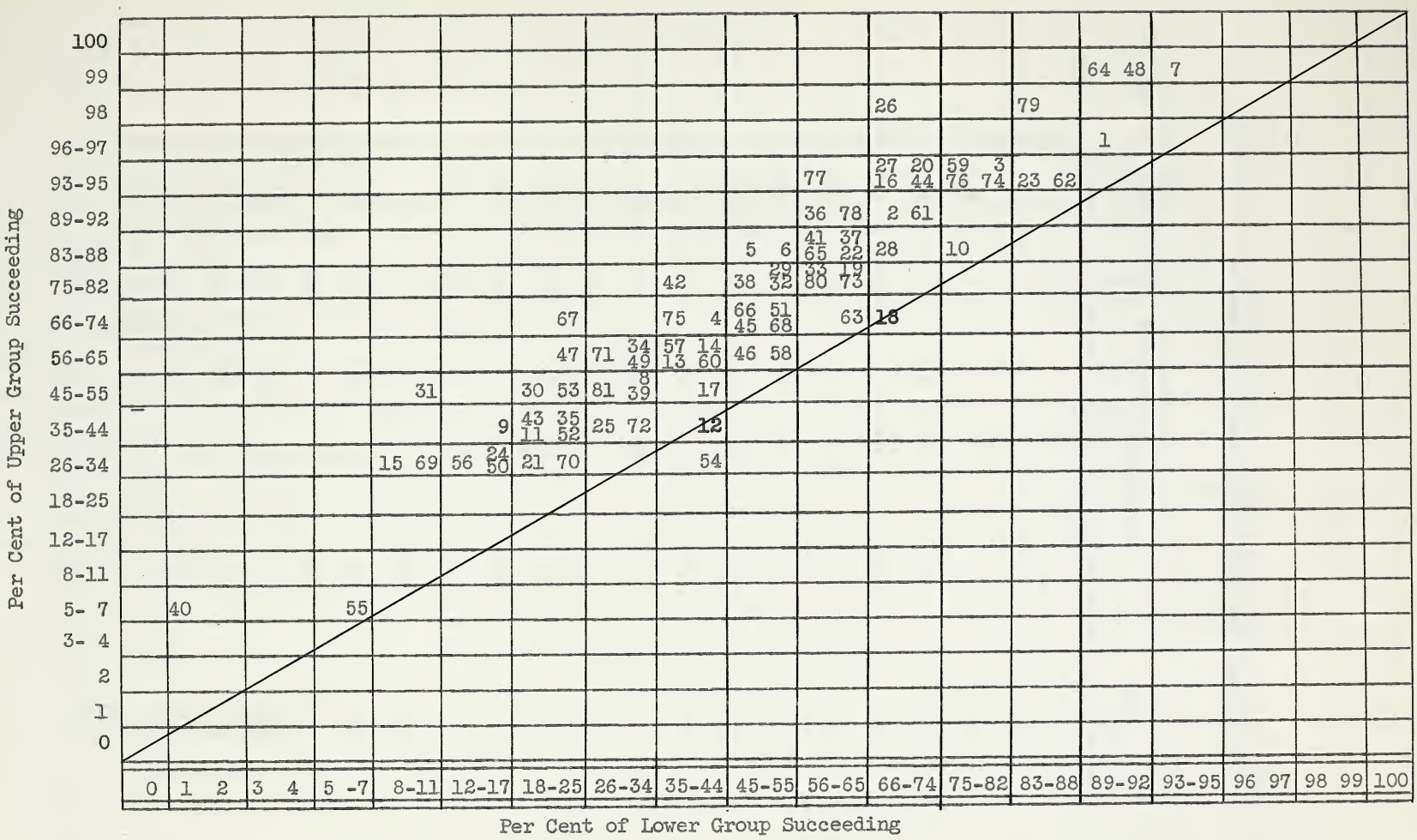


Fig. 3. Chart to facilitate use of Kelley Technique.

TABLE XIV. SUMMARY OF DATA ON INDIVIDUAL ITEMS FOR STANDARD ERROR OF THE DIFFERENCE BETWEEN TWO PERCENTAGES

Item No.	Highest 27%		Lowest 27%		$P_1 - P_2$	SD_{diff}	$\frac{P_1 - P_2}{SD_{diff}}$
	N_1	P_1	N_2	P_2			
1	105	97	99	92	5	3.1	1.61
2	99	92	74	69	23	5.16	4.46
3	103	95	83	77	18	4.56	3.9
4	80	74	46	43	31	6.4	4.8
5	90	83	52	48	35	6.01	5.8
6	93	86	55	51	35	5.9	5.9
7	107	99	103	95	4	2.3	1.7
8	54	50	36	32	18	6.6	2.7
9	46	43	18	17	26	5.98	4.3
10	94	87	82	76	11	5.23	2.1
11	43	40	20	19	21	6.04	3.48
12	40	37	39	36	1	6.5	0.15
13	70	65	42	39	26	6.5	4.0
14	61	56	43	40	16	6.7	2.4
15	33	31	12	11	20	5.37	3.73
16	100	93	77	71	22	5.0	4.4
17	58	54	41	38	16	6.7	2.4
18	78	72	72	67	5	6.2	.81
19	81	75	61	56	19	6.3	3.0
20	102	94	79	73	21	4.8	4.3
21	36	32	26	24	8	6.1	1.31
22	93	86	70	65	21	5.7	3.7

TABLE XIV - (Continued)

23	101	94	92	85	9	4.1	2.2
24	37	34	15	14	20	5.65	3.54
25	41	38	34	31	7	6.4	1.1
26	106	98	73	68	30	4.7	6.3
27	100	93	78	72	21	4.9	4.3
28	94	87	77	71	16	5.4	3.0
29	83	77	51	47	30	6.3	4.7
30	51	47	20	19	28	6.1	4.6
31	58	54	11	10	44	5.6	7.86
32	87	81	56	52	29	6.1	4.7
33	95	88	61	56	32	5.7	5.6
34	62	57	33	31	26	6.5	4.0
35	39	36	24	22	14	6.1	2.3
36	97	90	65	60	30	5.5	5.4
37	94	87	60	56	31	5.8	5.3
38	85	79	49	45	34	6.2	5.5
39	54	50	34	31	19	6.5	2.9
40	8	7	1	1	6	2.6	2.3
41	92	85	69	64	21	5.8	3.6
42	87	81	45	42	39	6.1	6.4
43	38	35	22	20	15	5.99	2.5
44	102	94	73	68	26	5.04	5.1
45	80	74	51	47	27	6.4	4.2

TABLE XIV - (Continued)

46	69	64	55	51	13	6.67	1.9
47	61	56	20	19	37	6.1	6.06
48	107	99	98	91	8	2.9	2.7
49	65	60	33	31	29	6.5	4.8
50	34	31	14	13	18	5.5	3.27
51	73	68	53	49	19	6.59	2.9
52	44	41	24	22	19	6.2	3.06
53	50	46	26	24	22	6.3	3.5
54	30	28	38	35	-7	6.3	-1.1
55	10	9	8	7	2	3.7	.54
56	31	29	18	17	12	5.7	2.1
57	61	56	41	38	18	6.7	2.7
58	64	59	55	51	8	6.75	1.18
59	101	94	75	69	25	5.0	5.0
60	69	64	45	42	22	6.6	3.3
61	96	89	78	72	17	5.3	3.2
62	102	94	95	88	6	3.9	1.5
63	80	74	64	59	15	6.3	2.4
64	107	99	97	90	9	3.04	2.9
65	91	84	65	60	24	5.9	4.1
66	80	74	57	53	21	6.4	3.3
67	76	70	26	24	46	6.1	7.5
68	77	71	59	55	16	6.5	2.5
69	35	32	10	9	23	5.2	4.4

TABLE XIV -(Concluded)

70	35	32	25	23	9	6.0	1.5
71	70	65	28	26	39	6.2	6.3
72	47	44	31	29	15	6.4	2.3
73	89	82	65	60	22	5.9	3.7
74	103	95	85	79	16	4.4	3.6
75	71	66	48	44	22	6.6	3.3
76	103	95	88	81	14	4.3	3.3
77	100	93	67	62	31	5.3	5.8
78	96	89	69	64	25	5.5	4.5
79	106	98	95	88	10	3.4	3.0
80	89	82	66	61	21	5.9	3.5
81	52	48	35	32	16	6.6	2.4

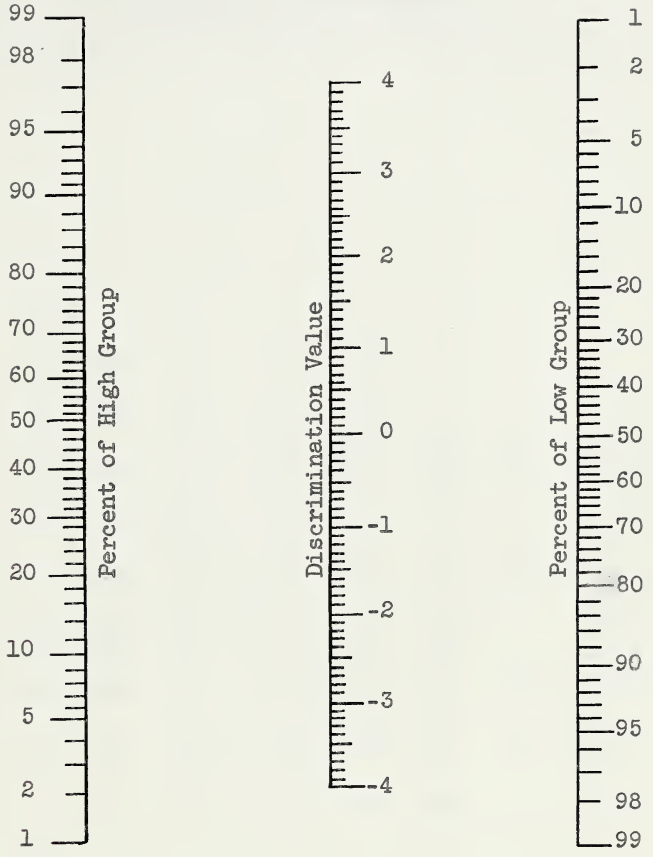


Fig. 4. Lawshe's Nomograph for Assigning Discrimination (D) Values to Test Questions.

C. H. Lawshe Jr., Principles of Personnel Testing, New York, McGraw-Hill Book Co. Inc., 1948, p.187.

TABLE XV. FREQUENCY DISTRIBUTION OF D-VALUES OF ITEMS BY LAWSHE'S NOMOGRAPH

D-VALUE	TABULATION (ITEM NUMBERS)										f
1.7	26										1
1.6											
1.5	31										
1.4											
1.3	67	77									2
1.2	4	6	42	44	59	64					6
1.1	2	5	33	36	47	71					6
1.0	3	16	20	27	37	38	48	78	79		9
0.9	29	32	69	74	76						5
0.8	7	9	15	22	30	40	41	45	49	65	10
0.7	11	13	24	34	61	73	80				7
0.6	19	23	28	50	52	60	75	53	66		9
0.5	1	8	10	39	43	51	57	63	68	17	10
0.4	14	46	56	62	72	81	35				7
0.3	21	70									2
0.2	18	25	58								3
0.1	12	55									2
0.0											
-0.1	54										1
										TOTAL	81

TABLE XVI. CALCULATION OF THE BI-SERIAL "R" BETWEEN TOTAL SCORES ON THE TEST AND THE RESPONSES TO A SINGLE ITEM ON THE TEST (ITEM NO. 21)

Scores on Test	Responses to Item 21		f
	Correct	Incorrect	
65 - 69	1	4	5
60 - 64	6	17	23
55 - 59	14	19	33
50 - 54	20	54	74
45 - 49	44	73	117
40 - 44	21	62	83
35 - 39	7	34	41
30 - 34	2	11	13
25 - 29	4	7	11
	p = 119	q = 281	N = 400

$M = 47.6$ Mean of 400 scores (Table I).
$\sigma = 8.2$ Standard Deviation of 400 scores (Table I).
$M_p = 47.55$ Mean of correct responses (N = 119).
$M_q = 46.44$ Mean of incorrect responses (N = 281).
$p = .30$ Proportion with correct response to Item 21.
$q = .70$ Proportion with incorrect response to Item 21.
$z = .348$ Height of ordinate separating 30% from 70% in a normal distribution (Table 54, Garrett).

$$r_{bis} = \frac{M_p - M_q}{\sigma} \times \frac{p\,q}{z}$$

$$= \frac{47.55 - 46.44}{8.2} \times \frac{.30 \times .70}{.348}$$

$$= \frac{1.11}{8.2} \times \frac{.21}{.348}$$

$$= .08$$

$$\sigma_{r_{bis}} = \frac{\left(\frac{\sqrt{pq}}{3} - r^2_{bis}\right)}{\sqrt{N}}$$

$$= \frac{\frac{\sqrt{.21}}{.348} - (.08)^2}{\sqrt{400}}$$

$$= .06$$

C.R. = .08 ÷ .06 = 1.33. From Table 29, Garrett, we find that this critical ratio is not significant at the 5% level. (At the 5% level C.R. = 1.97 when (N-1)= 399.)

TABLE XVII. SUMMARY OF BI-SERIAL R CALCULATIONS ON ITEMS 21, 35, 43, 52 AND 70

Scores on Test	Item 21			Item 35			Item 43			Item 52			Item 70		
	p	q	f	p	q	f	p	q	f	p	q	f	p	q	f
65 - 69	1	4	.5	5	0	5	3	2	5	4	1	5	3	2	5
60 - 64	6	17	23	11	12	23	11	12	23	12	11	23	10	13	23
55 - 59	14	19	33	17	16	33	14	19	33	9	24	33	11	22	33
50 - 54	20	54	74	25	49	74	22	52	74	29	45	74	23	51	74
45 - 49	44	73	117	29	88	117	34	83	117	37	80	117	26	91	117
40 - 44	21	62	83	21	62	83	12	71	83	22	61	83	19	64	83
35 - 39	7	34	41	7	34	41	6	35	41	11	30	41	6	35	41
30 - 34	2	11	13	2	11	13	2	11	13	2	11	13	5	8	13
25 - 29	4	7	11	4	7	11	2	9	11	2	9	11	0	11	11
	119	281	400	121	279	400	106	294	400	128	272	400	103	297	400
	M = 47.6 σ = 8.2 Mp = 47.55 Mq = 46.44 p = .30 q = .70 z = .348			M = 47.6 σ = 8.2 Mp = 49.25 Mq = 45.65 p = .30 q = .70 z = .348			M = 47.6 σ = 8.2 Mp = 49.7 Mq = 45.7 p = .265 q = .735 z = .328			M = 47.6 σ = 8.2 Mp = 48.6 Mq = 45.9 p = .32 q = .68 z = .358			M = 47.6 σ = 8.2 Mp = 49 Mq = 46 p = .26 q = .74 z = .324		
Bi-serial Validity Index	r_{bis} = .06			r_{bis} = .20			r_{bis} = .21			r_{bis} = .15			r_{bis} = .15		
Davis Index	26.5			31			30			36.5			25		
Flanagan	.09			.16			.19			.21			.12		

TABLE XVIII. ITEMS ON THE PICTURE TEST IN SCIENCE WHICH FAIL TO DISCRIMINATE AS SHOWN BY A COMBINED SUMMARY OF FIVE METHODS

Davis Index	Flanagan Coefficient	Kelly	Standard Error of Differences Between %'s		Lawshe's Nomograph Values	
			5%	1%	.3	.5
1	1	1	1	1		1
7	7		7	7		
8	8	8				8
10	10	10		10		10
12	12	12	12	12	12	12
14	14	14		14		14
17	17	17		17		17
18	18	18	18	18	18	18
19	19	19				
	21	21	21	21	21	21
25	25	25	25	25	25	25
28	28	28				
	35	35		35		35
39	39	39				39
40	40			40		
41	41					
	43	43		43		43
46	46	46	46	46		46
51	51	51				51

TABLE XVIII - (Concluded)

	52	52				
54	54	54	54	54	54	54
55	55	55	55	55	55	55
56	56	56		56		56
57	57	57				57
58	58	58	58	58	58	58
	60	60				
61	61					
62	62	62	62	62		62
63	63	63		63		63
66	66					
68	68	68		68		68
	70	70	70	70	70	70
72	72	72		72		72
73	73					
75	75	75				
76	76					
80	80					
81	81	81		81		81

CHAPTER VIII

CONCLUSIONS

1. The validity of the "Picture Test in Science" as determined by correlating 100 Grade VII test scores against teachers' estimates of pupil ability was .27. Although the test is only in an experimental form this is a very low validity. This may be due to the fact that the criterion used for obtaining the coefficient was unreliable, but nevertheless the test needs to be improved before it is printed in its final form. The validity of the test can be improved by discarding items of low or negative validity and by revising others so that they will be more valid.
2. The three reliability coefficients obtained on the "Picture Test in Science" reveal important information. A correlation of .78 was secured from the Split-Half Method corrected by the Spearman-Brown formula and an identical result was secured from the Analysis of Variance Method. These two methods are based on entirely different procedures and this would seem to indicate that the reliability coefficient obtained is an accurate one. Furthermore, this receives support from the third method used for reliability. A coefficient of .72 was yielded by the Method of Rational Equivalence. At first glance this would appear to be in disagreement with the other methods, but it must be remembered as was stated in Chapter IV that the approximation formula used for this method always underestimates the reliability of a test as found by the Split-Half technique. If a test is being used to differentiate among the

individuals in the group being tested its reliability should be .90 or more. Therefore the obtained reliability coefficient of .78 for the "Picture Test in Science" is low. The reliability of a test can be increased by the expedient of lengthening the test, by excluding items that tend to lower the reliability, or by replacing certain items by others that will increase reliability. The first method suggested is out of the question because the test is too long in its present form therefore it would be necessary to try the other suggested methods.

3. The Davis Method is the best item-analysis procedure because it provides a set of difficulty indices in addition to the set of discrimination indices. No other method does this. Also, it is the only method used in this study which makes a correction for chance. The Davis Method appears to have a serious weakness as shown by the bi-serial r calculations on the five doubtful items. The use of the special table recommended when the proportion of successes is negative gives a spuriously high validity index for the item. The bi-serial r coefficient is recognized as producing the most reliable validity coefficient and its use on the five doubtful items indicated that they possessed low validity. This result was further borne out by the fact that each of the other methods used in this study supported this conclusion.

4. It is wasteful to use item-analysis techniques to weed out items that are obviously defective to a subject-matter expert. This seems to be the case as far as the "Picture Test in Science" is concerned. A more careful scrutiny of the test items should have

been made before the test was printed in its present form and better results would have been secured. When a large proportion of the items are defective the criterion, which is the total score on the test, will include extraneous and unwanted functions.

5. It is not worthwhile to use the Analysis of Variance procedure to secure only a reliability coefficient as was done in this study. Although it provided a valuable result in connection with reliability the amount of computation required to obtain it is entirely out of proportion to the results obtained. A useful study could be carried out using Analysis of Variance as the basic technique for item-analysis. The writer of this study has discovered than an Analysis of Variance table provides an ideal set-up for test data. In addition to providing a reliability coefficient, the table may be used for the following procedures:
   (a) the high and low 27% of the class required for most item-analysis methods can be extracted from the table quite easily;
   (b) a counting device can be constructed to sub-divide test scores into odd-even classification for split-half method directly from the table instead of going back to the test papers to secure this information;
   (c) the laborious tabulation of responses required for each item when making use of the bi-serial r can be simplified by use of this chart.
6. The "Picture Test in Science" was top-heavy with easy items. Using the 85% pass or fail dividing line 10 items (Nos. 1, 3, 7, 23, 48, 62, 64, 74, 76, and 79) were too easy whereas only 2 items (Nos. 40 and 55) were too hard. These items should be eliminated from the

next edition of the test. It would be permissible to retain a few of the easy items to use as lead-on questions at the beginning of the test. If this procedure is followed the choice would have to be made from items 3, 48, 64, 74, or 79 because these are the only easy items that are valid.

7. The following items have low discriminating power as shown by the summary of five item-analysis methods used in this study and they should be eliminated from the revised test: 1, 7, 8, 10, 12, 14, 17, 18, 19, 21, 23, 25, 28, 35, 39, 40, 41, 43, 46, 51, 52, 54, 55, 56, 57, 58, 61, 62, 63, 68, 70, 72, 73, 75, 76, 80, and 81.
8. The remaining items should be retained because item-analysis procedures show that they possess sufficient discriminating power and also because they are within the required range of difficulty. On a revised test these items would be arranged in order of difficulty from the easiest to the hardest. This arrangement would be: 20, 26, 27, 16, 59, 44, 2, 77, 78, 36, 22, 65, 37, 33, 6, 32, 5, 66, 38, 29, 45, 42, 4, 13, 60, 67, 71, 49, 34, 47, 31, 53, 30, 9, 11, 24, 69, 50, and 15. This arrangement is based on difficulty indices secured from the Davis item-analysis method. Thus the "Picture Test in Science" would be reduced from its present experimental form of 81 questions to a revised form containing 39 valid items arranged in order of difficulty.

BIBLIOGRAPHY

Clark, E.L. "A Method of Evaluating the Units of a Test". Journal of Educational Psychology XIX: 263-65, April 1928.

Cronbach, Lee J. Essentials of Psychological Testing. New York, Harper, 1949. 475 p.

Davis, Frederick B. Item - Analysis Data, Their Computation, Interpretation and Use in Test Construction. Harvard Education Papers, No. 2. Cambridge, Mass. Graduate School of Education, Harvard University. 1946. 42 p.

Edwards, Allen L. Statistical Analysis. New York, Rinehart and Company, Inc., 1947. 360 p.

Fisher, Ronald A. The Design of Experiments. New York, Hofner Publishing Company, Inc., 1949. 240 p.

Fisher, Ronald A. Statistical Methods for Research Workers. London, Oliver and Boyd, 1938. p 203.

Froelich, G. J. "A Simple Index of Test Reliability". Journal of Education Psychology XXXII: 381-85, 1941.

Flanagan, J.C. "General Considerations in the Selection of Test Items". Journal of Educational Psychology XXX: 674-80, 1939

Garrett, Henry E. Statistics in Psychology and Education. New York, Longmans, Green and Company, 1947. 487 p.

Hawkes, H.E., Lindquist, E.F., and Mann, C.R. The Construction and Use of Achievement Examinations. Boston, Houghton Mifflin Company, 1936, 491 p.

Horst, A.P. "The Difficulty of a Multiple-Choice Test Item". Journal of Educational Psychology XXIV: 30, March, 1933.

Jackson, Robert W.B. Application of the Analysis of Variance and Covariance Method to Educational Problems. Bulletin No. 11, Department of Educational Research Toronto, University of Toronto, 1940. 103 p.

Jackson, Robert W. B. and Ferguson, George A. Studies on the Reliability of Tests. Bulletin No. 12, Department of Educational Research. Toronto, University of Toronto, 1941, 132 p.

Kelley, T.L. "Selection of Upper and Lower Groups". Journal of Educational Psychology XXX: 17-24, 1939.

Kelley, T.L. Statistical Method. New York, Macmillan Company, 1924. p. 373-85.

Lawshe, Charles H. "Item - Analysis Procedures". Principles of Personnel Testing. New York, McGraw Hill Book Company Inc., 1948. p. 183-91.

Lentz, T.F. "Evaluation of Methods of Evaluating Test Items". Journal of Educational Psychology XXIII: 344-50, May 1932.

Lindquist, E.F. Statistical Analysis in Educational Research. Boston, Houghton Mifflin Company, 1940. 266 p.

Long, John A. and Sandiford, Peter and others. The Validation of Test Items. Bulletin No. 3 of the Department of Educational Research, University of Toronto, 1935, 126 p.

McCall, William A. Measurement. New York, The Macmillan Company, 1939. 535 p.

McNemar, Quinn. Psychological Statistics. New York, John Wiley and Sons, Inc., 1949. 364 p.

Micheels, William J. and Karnes, M. Ray. "Analyzing and Improving Tests". Measuring Educational Achievement. New York, McGraw Hill Book Company Inc., 1950. p. 454-84.

Monroe, W.S. Encyclopedia of Educational Research. American Educational Research Association. New York, Macmillan, 1950. p. 1016-17 and p. 1461-77.

Richardson, M.W. and Kuder, G.F. "The Calculation of Test Reliability Coefficients Based on the Method of Rational Equivalence". Journal of Educational Psychology XXX: 681-87, 1939.

Ross, C.C. Measurement in Today's Schools. New York, Prentice-Hall Inc., 1947. 551 p.

Symonds, P.M. "Choice of Items for a Test on the Basis of Difficulty". Journal of Educational Psychology XX: 481-93, October, 1929.

Symonds, P.M. "Factors Influencing Test Reliability". Journal of Educational Psychology XIX: 73-87, 1928.

Symonds, P.M. Measurement in Secondary Education. New York, The Macmillan Company, 1927, 588 p.

Weitzman, Ellis and McNamara, Walter J. "Test Analysis". Constructing Classroom Examinations. Chicago, Science Research Associates, 1949. p 94-107.

FOOTNOTES

1. Monroe, W.S. Encyclopedia of Educational Research. Revised Edition. New York, Macmillan, 1950. p. 1471

2. Ibid. p. 1461

3. Ibid.

4. Ibid.

5. Weitzman, Ellis and McNamara, Walter J. Constructing Classroom Examinations. Chicago, Science Research Associates, 1949. p. 3.

6. Symonds, P.M. "Factors influencing Test Reliability." Journal of Educational Psychology. XIX: 73-87, 1928.

7. Long, John A. and Sandiford, Peter and others. The Validation of Test Items. Bulletin No. 3 of the Department of Educational Research, University of Toronto, 1935, p. 7.

8. Ibid. p. 10

9. Ibid. p. 16.

10. Ibid. p. 18.

11. Ibid. p. 30.

12. Clarke, E.L. "A Method of Evaluating the Units of a Test." Journal of Educational Research. XIX, 1928. p. 263.

13. Long, John A. and Sandiford, Peter, op. cit. p. 43.

14. Richardson, M.W. And Kuder, G.F. "The Calculation of Test Reliability Coefficients Based on the Method of Rational Equivalence." Journal of Educational Psychology. XXX, 1939. p. 681.

15. Froelich, G.J. "A Simple Index of Test Reliability." Journal of Educational Psychology. XXXII, 1941 p. 382

16. Ibid. p. 385

17. Micheels, William J. and Karnes, M. Ray. "Analyzing and Improving Tests." Measuring Educational Achievement. New York, McGraw Hill Book Company Inc., 1950. p. 472.

18. Monroe, W.S. op. cit. p. 1471

19. Davis, Frederick B. Item-Analysis Data, Their Computation, Interpretation and Use in Test Construction. Harvard Education Papers, No. 2. Cambridge, Mass. Graduate School of Education, Harvard University. 1946. p.1

20. Ibid. p. 3

21. Horst, A.P. "The Difficulty of a Multiple-Choice Item." Journal of Educational Psychology. XXIV, 1933. p. 30.

22. Kelley, T.L. Statistical Method. New York, Macmillan Company, 1924. p. 373.

23. Kelley, T.L. "Selection of Upper and Lower Groups." Journal of Educational Psychology. XXX, 1939. p. 17

24. Flanagan, J.C. "General Considerations in the Selection of Test Items." Journal of Educational Psychology. XXX, 1939. p.674

25. Davis, Frederick B. op. cit. p. 15

26. Flanagan, J.C. op. cit. p. 678

27. Long, John A. and Sandiford, Peter. op. cit. p. 34

28. Garrett, Henry E. Statistics in Psychology and Education. New York, Longmans, Green and Company, 1947. p. 190.

29. Lawshe, Charles H. "Item-Analysis Procedures." Principles of Personnel Testing. New York, McGraw Hill Book Company Inc., 1948. p. 187.

30. Ibid. p. 190.

Picture Test in Science

FOR GRADES FOUR, FIVE and SIX

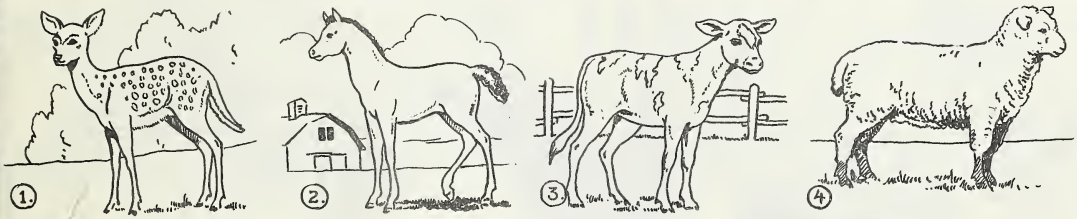
Experimental Edition Form A

Name: Grade: Date:

DIRECTIONS TO PUPILS: Here is a test arranged to find out what you have learned about many of the interesting things in science. For each question there are several answers illustrated with pictures. In each case, decide which picture is the correct answer and then write its number in the bracket to the right of the question.

SAMPLE A

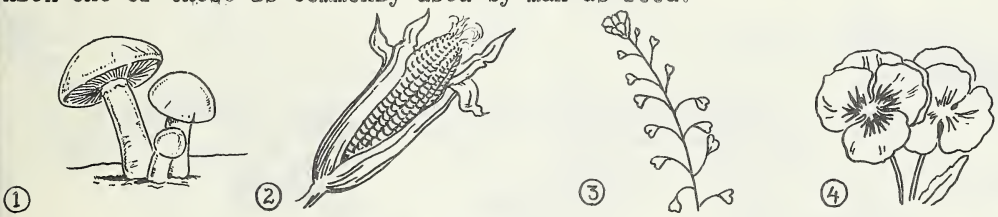
Which one is NOT like the others?



(4)

SAMPLE B

Which one of these is commonly used by man as food?



(2)

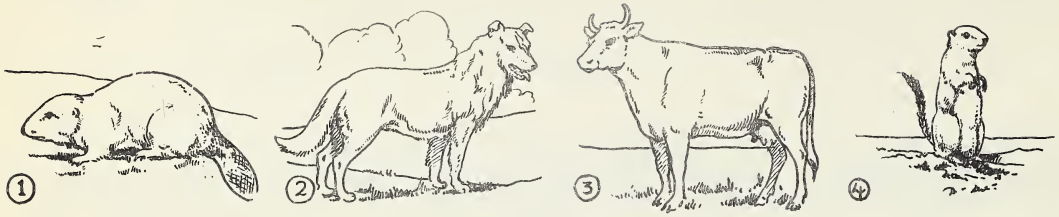
In sample A, the only one which is not like the others is the sheep. The others are all baby animals. Therefore, its number 4 has been placed in the space at the right.

In sample B, corn is of course used commonly by man as food so its number 2 is written in the space at the right.

Now, when your teacher tells you to begin, do all the questions in the same way. Work carefully but do not waste time over any one question. When you finish one page go right on to the next.

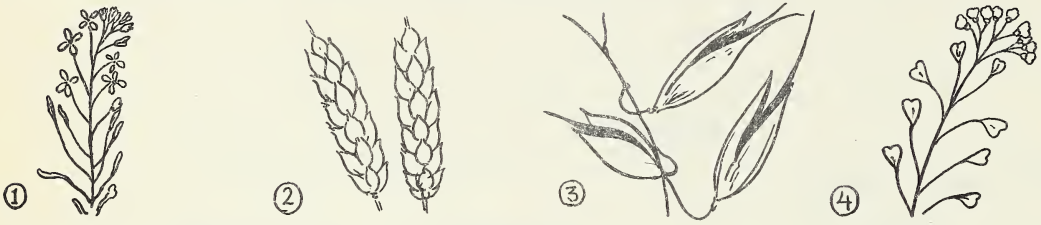
PART 1 - LIVING THINGS

1. Which of the following animals does man use for food ?



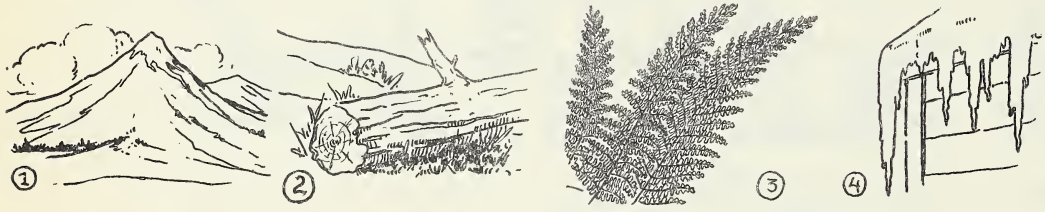
( )

2. Which plant is of value to man as food ?



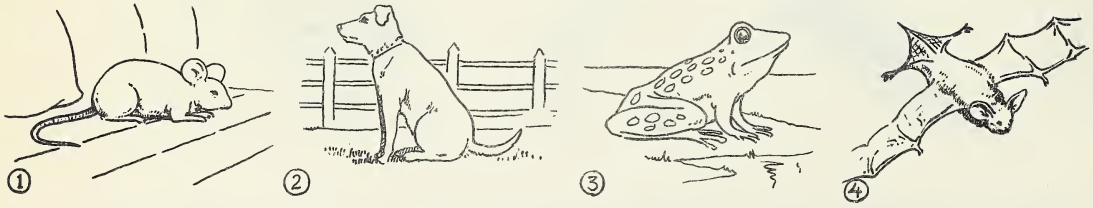
( )

3. Which of the following is a living thing ?



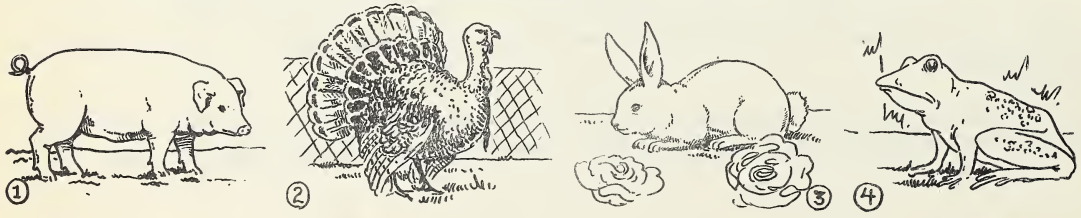
( )

4. Which of these is an enemy of man ?



( )

5. Which one of the following does more good than harm in your garden ?

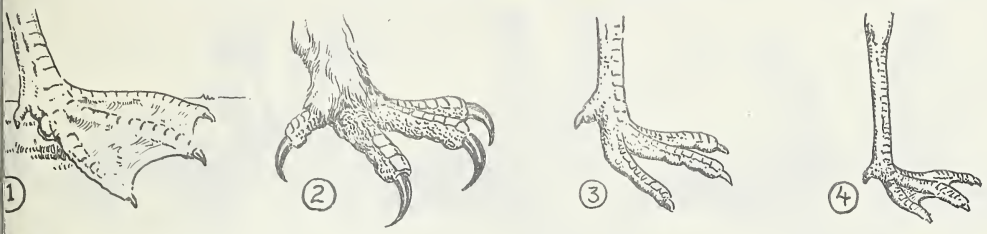


( )

Which Beak is best suited to a flesh-eating bird

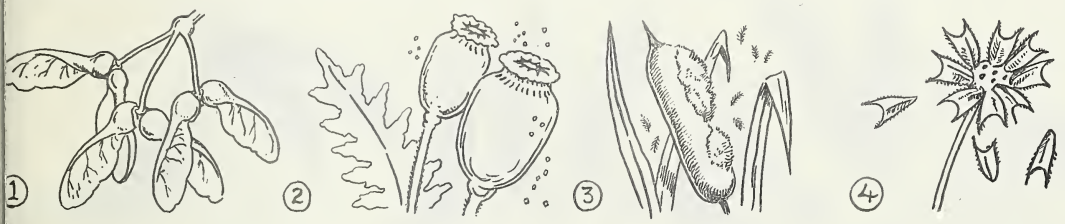
( )

Which of these feet is best adapted for swimming?



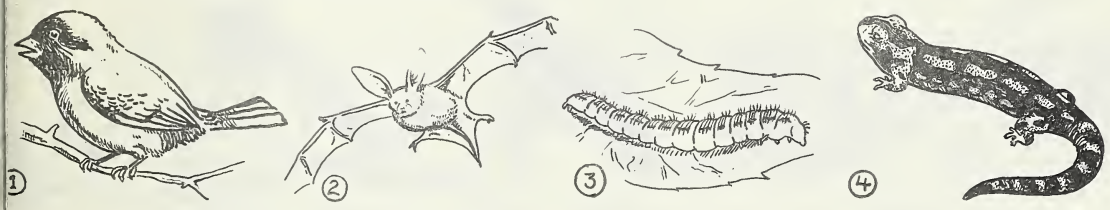
( )

Which one of these plants makes the most use of the wind to scatter its seeds?



( )

Which of the following is a mammal?

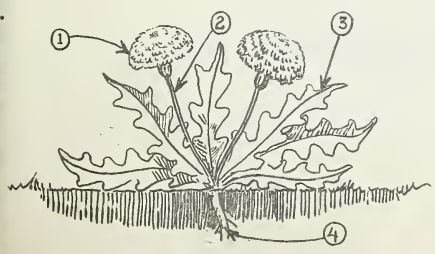


( )

. Which of these plants grows from a bulb?

( )

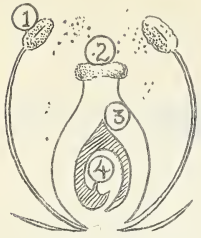
..



In which part of this plant is most of its food manufactured?

( )

12. In the flower below where are the seeds formed?

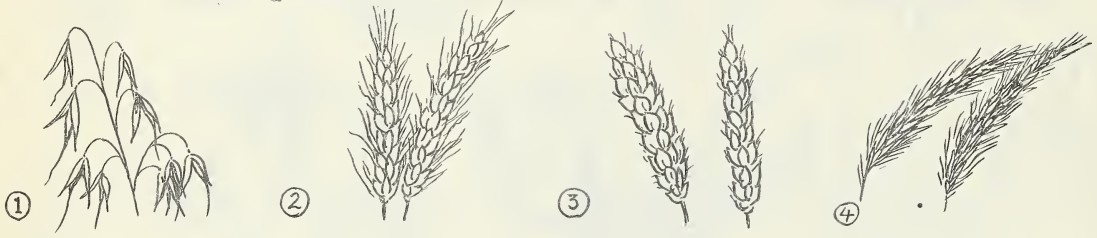


( )

13. Which of the following common weeds is Ball Mustard?

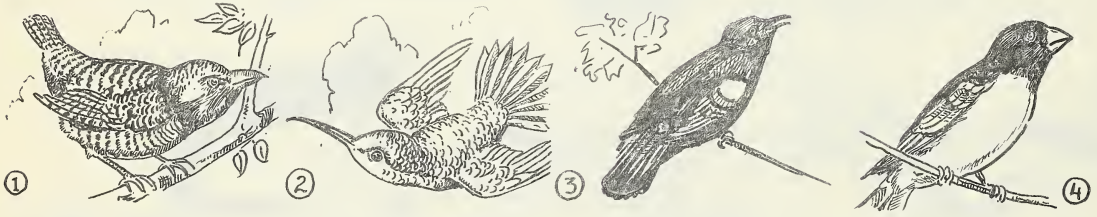
( )

14. Which of these grains is Wheat ?



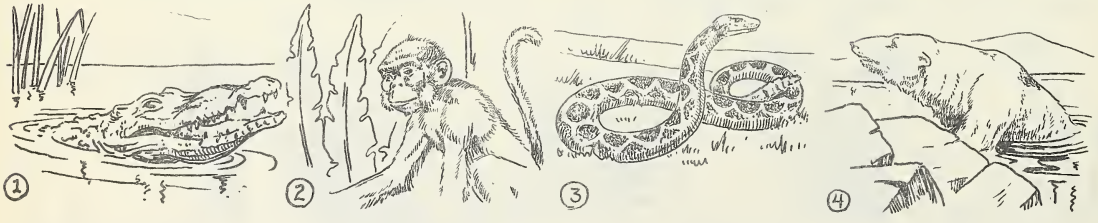
( )

15. Which of these birds is the grosbeak ?



( )

16. Which one of these animals would most likely be found in a swamp community?



( )

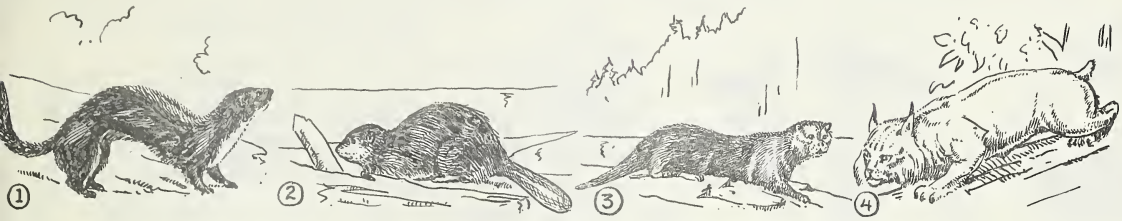
GO RIGHT ON TO THE NEXT PAGE

In each question below, find the one which is NOT like the others. Then put its number in the bracket at the right.

17.

( )

18.

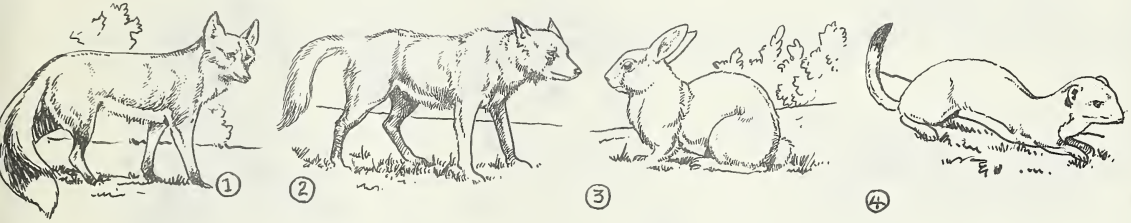


( )

19.

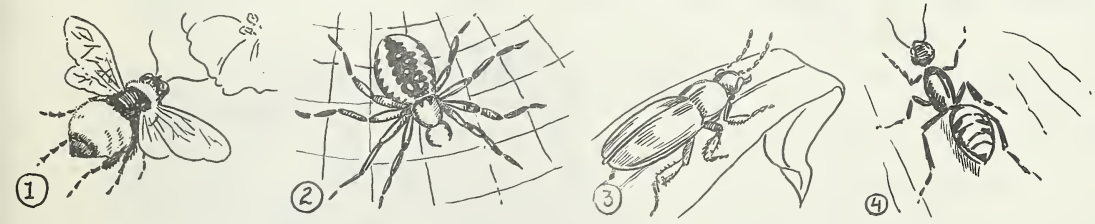
( )

20.



( )

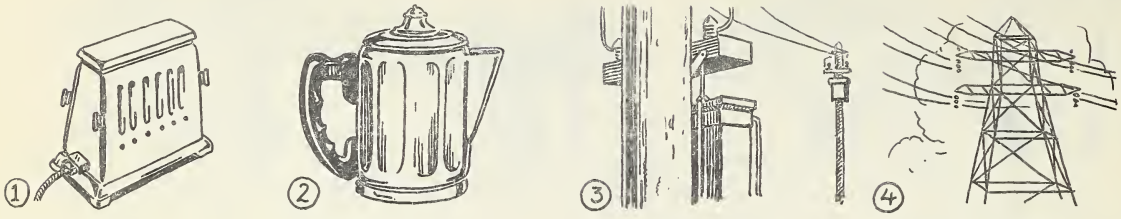
21.



( )

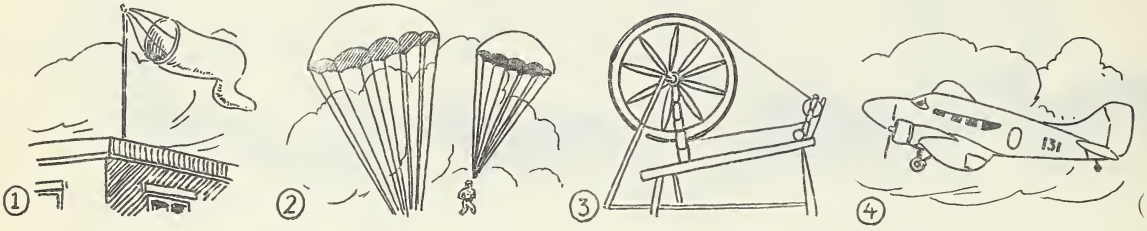
GO RIGHT ON TO THE NEXT PAGE

22. In each question below find the one which is NOT like the others. Then put its number in the bracket at the right.



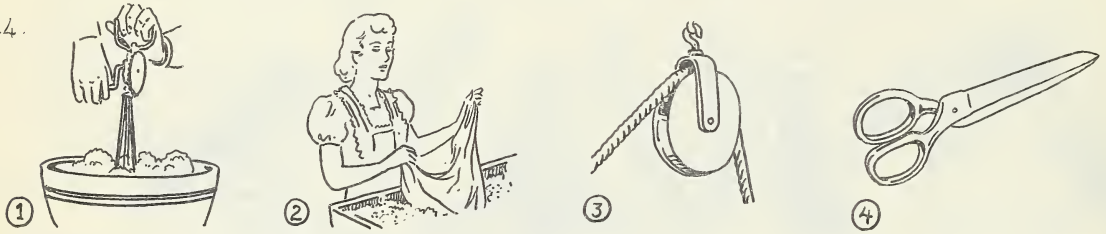
( )

23.



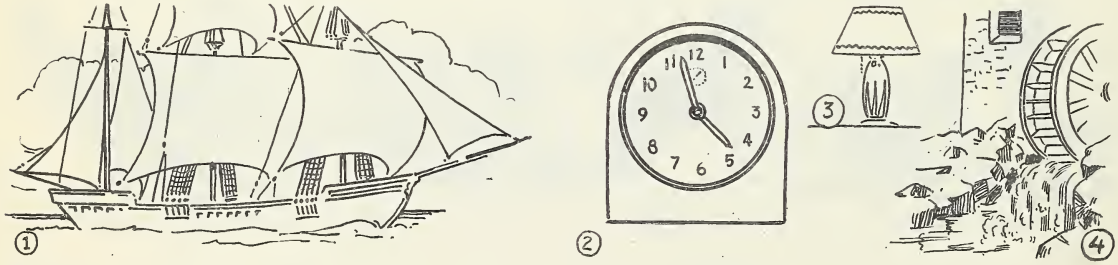
( )

24.



( )

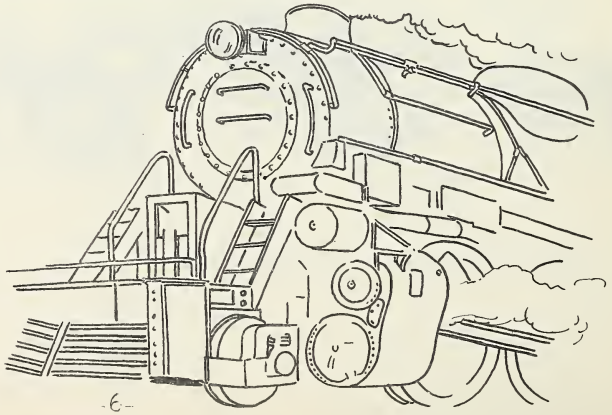
25. Which of the following illustrates the form of power longest used by man ?



( )

26. Which form of power is being used for moving this huge machine?

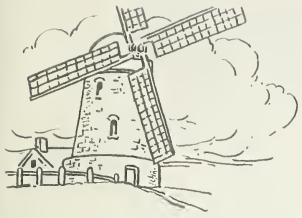
1. electricity
2. wind
3. steam
4. water



( )

GO RIGHT ON TO THE NEXT PAGE

27. Here is an illustration of how man gets work done by the use of Nature's forces. What force is being made use of ?



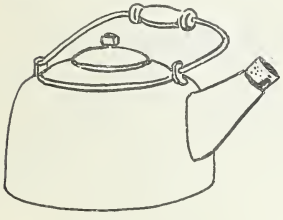
1. Electric Power
2. Water Power
3. Wind Power
4. Horse Power

( )

28. How many people are in the following two pictures? Write your answer in the answer bracket at the right.

( )

29. Imagine the spout of a kettle is securely plugged while the heat is kept on and the lid loose. What will happen?



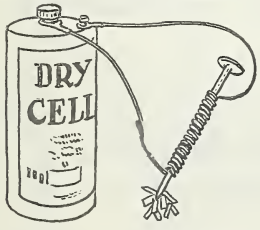
1. The kettle will boil dry.
2. The kettle will burst
3. The lid will lift up.
4. Nothing will happen.

30. Imagine that a piece of rubber balloon is fastened over the top of the bottle when it is full of steam. What will happen after it has cooled?

1. Nothing will happen.
2. The rubber will be forced off the top of the bottle.
3. The rubber over the opening will bulge upwards.
4. The rubber over the opening will bulge inward.

( )

31. Mark what will happen if I disconnect the wires from the battery.

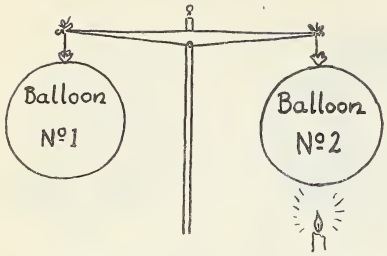


1. The pieces of metal will drop.
2. Nothing will happen.
3. The battery will go dead.
4. Sparks will fly from the nail.

( )

GO RIGHT ON TO THE NEXT PAGE

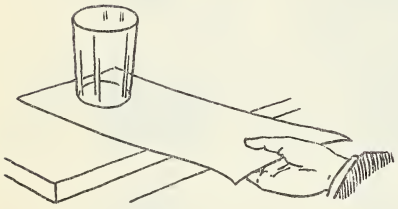
32. As Balloon No.2 fills with heated air from the candle mark what will happen.



1. Nothing will happen
2. The balloon will take fire.
3. Number 1 Balloon will drop lower.
4. Both balloons will rise.

( )

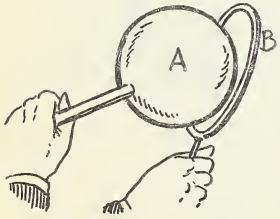
33. A sheet of smooth paper is to be jerked from under the tumbler. What will happen?



1. The tumbler will be jerked off the table.
2. The paper will not come from under the tumbler.
3. The tumbler will fall over.
4. The paper will come away leaving the glass upright.

( )

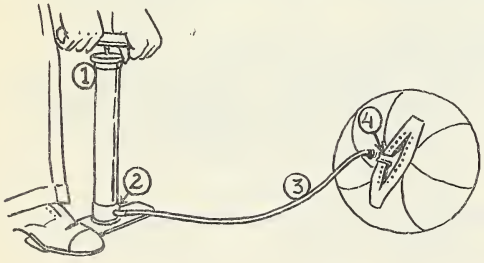
34. Metal ball "A" will nearly go through the metal ring "B". Which method would you use to get it through easily?



1. Pound it through with a hammer.
2. Heat the ring.
3. Heat the ball.
4. Make the ball smaller.

( )

35. Here is a football being pumped up. Which point becomes the hottest?



( )

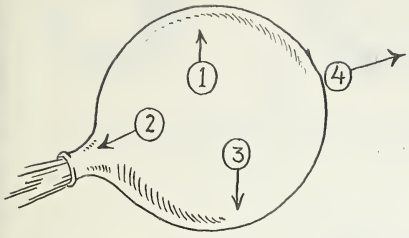
36. Below is a tumbler of ice water. Why will the tumbler get moist on the outside?

1. Water seeps through the glass.
2. The person who filled the glass made the outside of the glass wet.
3. Moisture condensed from the air.
4. The glass overflowed as the ice melted.

( )

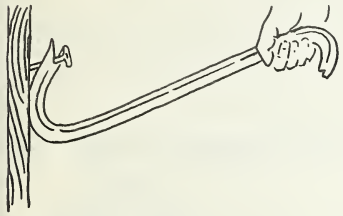
GO RIGHT ON TO THE NEXT PAGE

37. When the balloon opening is released in which direction will the balloon move?



( )

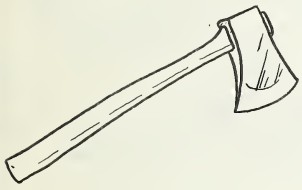
38. A nail-puller is better than a hammer for pulling out large nails. Why?



1. It is more convenient.
2. It holds the nail well.
3. It is stronger than a hammer.
4. It gives you more leverage than a hammer.

( )

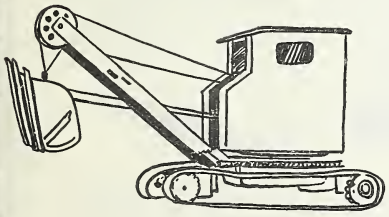
39. The axe is a machine called a wedge. It uses the principle of the:



1. Inclined plane
2. Pulley.
3. Lever.
4. Gear.

( )

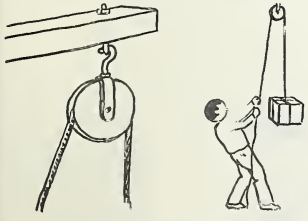
40. From where originally does the power come to run this machine?



1. Gasoline.
2. Wind.
3. Sun.
4. Water Power.

( )

41. In this picture a man is lifting a weight by means of a pulley. Why does the pulley make it easier?

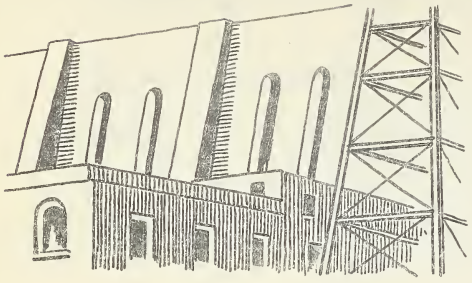


1. Less pull is needed.
2. the pulley is greased
3. The pulley is big.
4. Pulling downward is easier than lifting.

( )

GO RIGHT ON TO THE NEXT PAGE

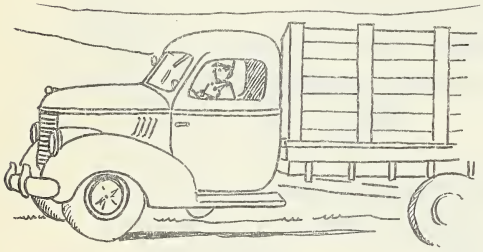
42. This plant produces electrical energy by the use of what form of power?



1. Steam.
2. Wind.
3. Water
4. Coal.

( )

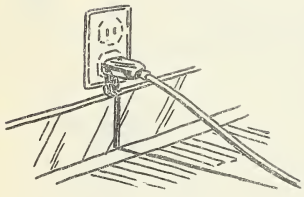
43. This truck most likely uses gasoline for power. Which of the following must it also have to make it move?



1. Tires.
2. Electricity.
3. Fenders.
4. Headlights.

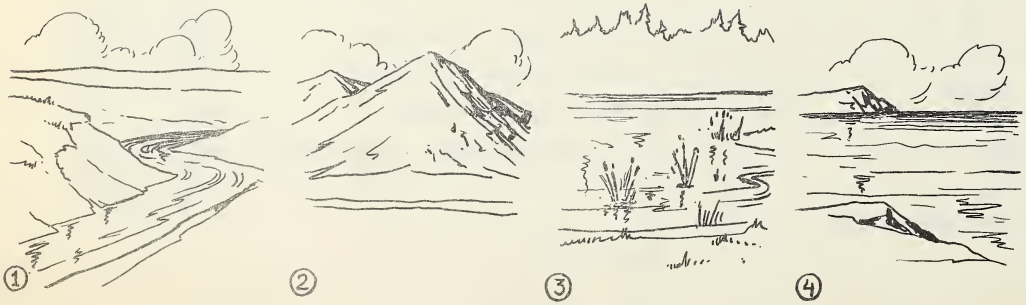
( )

44. Most of the wires conducting electricity are covered with rubber. Why?



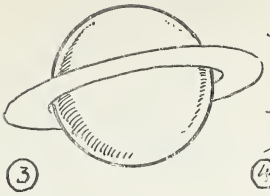
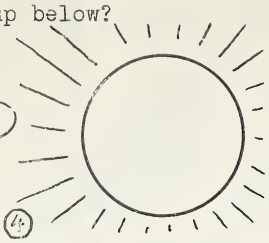
1. It looks better.
2. It helps carry the electric current.
3. It prevents escape of electricity.
4. It is cleaner. ( )

45. Which of the following does not belong in the group below?



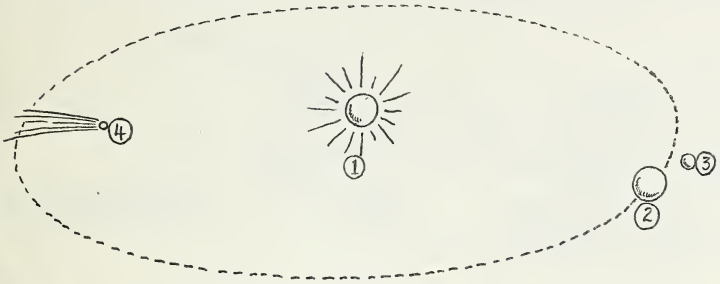
( )

46. Which of the following does not belong in the group below?

① ② ③ ④ ( )

47. In the diagram below, which is the moon?



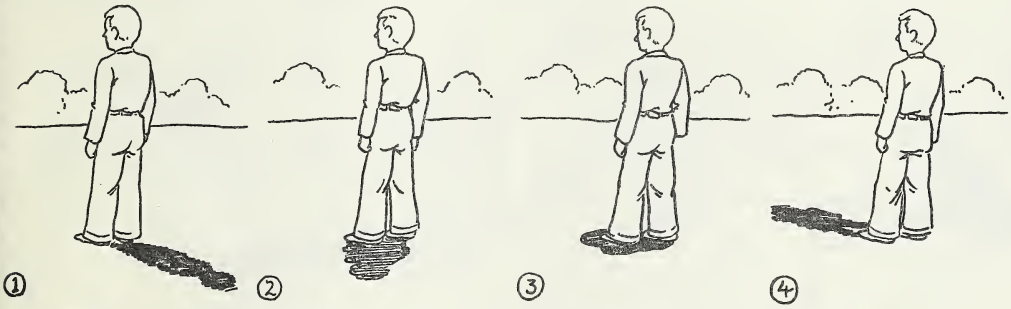
( )

48. Which instrument below is used for telling direction?

( )

① ② ③ ④

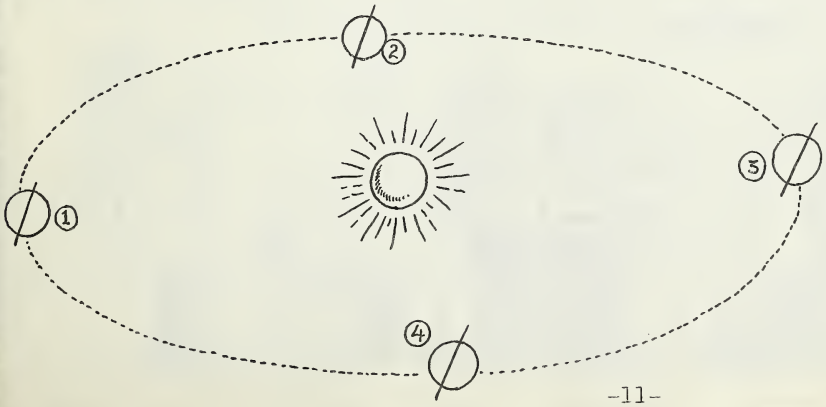
49. In each of the pictures below, the boy is facing east. By studying the shadows decide which picture represents evening.



( )

① ② ③ ④

50. Which is the position of the earth during our summer month .



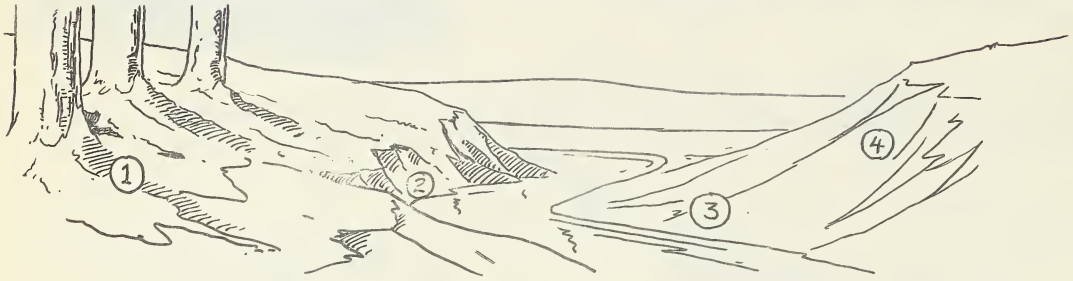
( )

GO RIGHT ON TO THE NEXT PAGE

51. Which is the North Star in the group?

4 1 3 2 ( )

52. Where will the snow be found latest in the spring?



( )

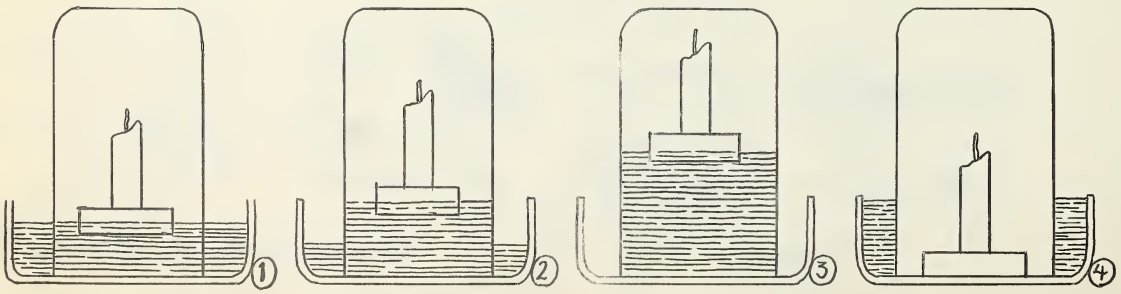
53. By studying the picture below decide where the most rain will fall.

( )

54. Where in the picture below would you expect the air to be rising in the daytime?

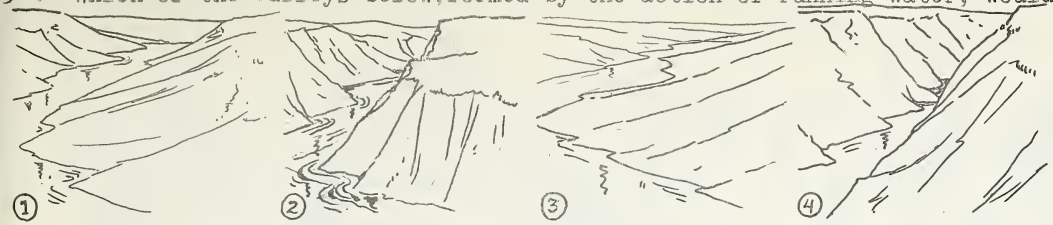
( )

55. A lighted candle resting on a cork float in a pan of water is covered by a glass jar. Which picture shows what happens when the candle finally goes out?



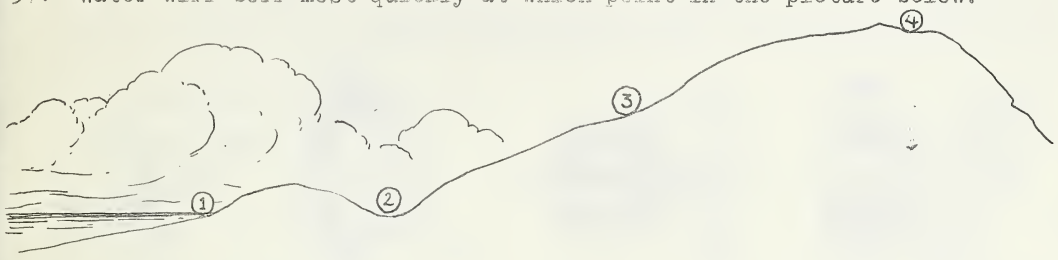
( )

56. Which of the valleys below, formed by the action of running water, would normally be the youngest?



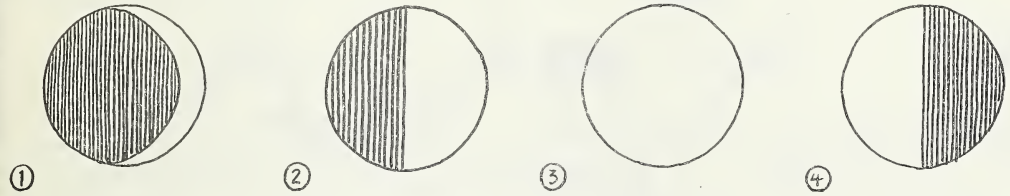
( )

57. Water will boil most quickly at which point in the picture below?



( )

58. Which of these diagrams shows the moon in the phase known as the "First Quarter"?

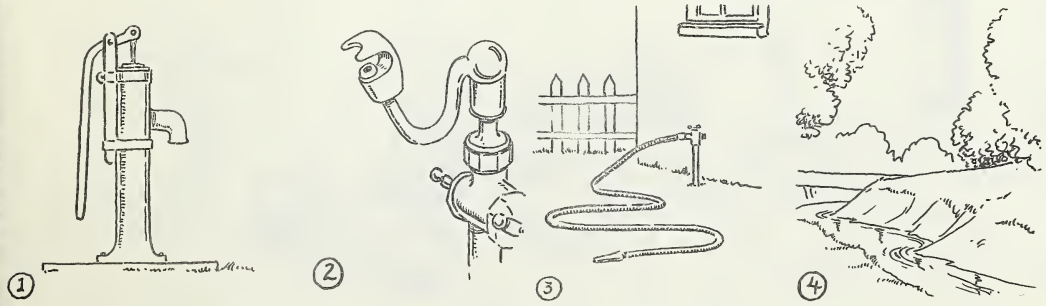


( )

59. These pictures illustrate different seasons. Which one illustrates autumn?

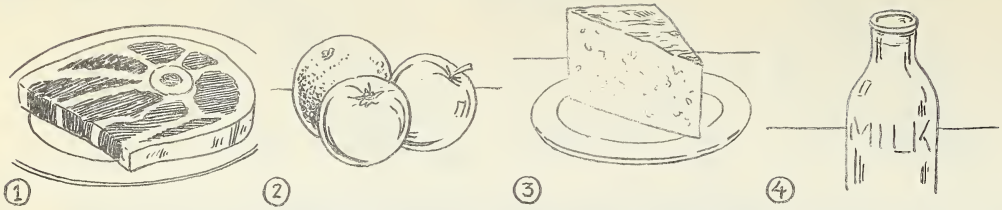
( )

60. Which of the following would be the safest place to get a drink of water ?



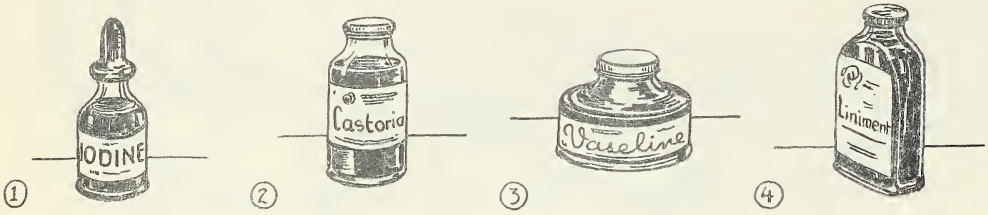
( )

61. Which is considered the "Perfect Food"?



① ② ③ ④ ( )

62. Which of these would you put on a cut to kill disease germs?



① ② ③ ④ ( )

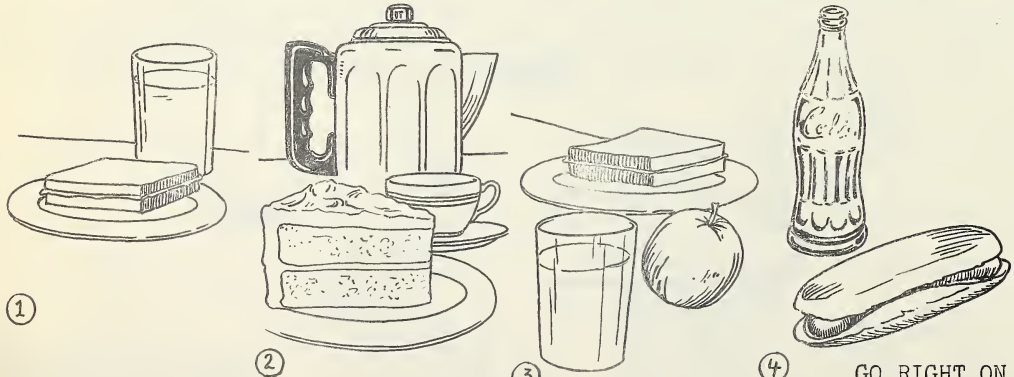
63. Which boy is taking good care of his eyes?

① ② ③ ④ ( )

64. Which boy is riding safely?

① ② ③ ④ ( )

65. Choose the best lunch for a growing boy or girl.



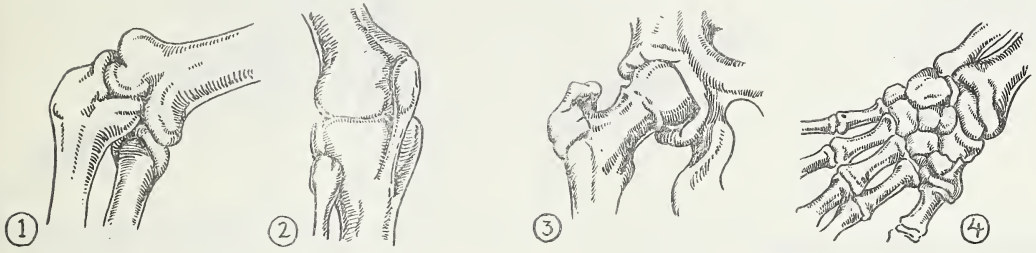
① ② ③ ④ ( )

GO RIGHT ON TO THE NEXT PAGE

66. Which picture shows the best way to prevent the spread of disease?

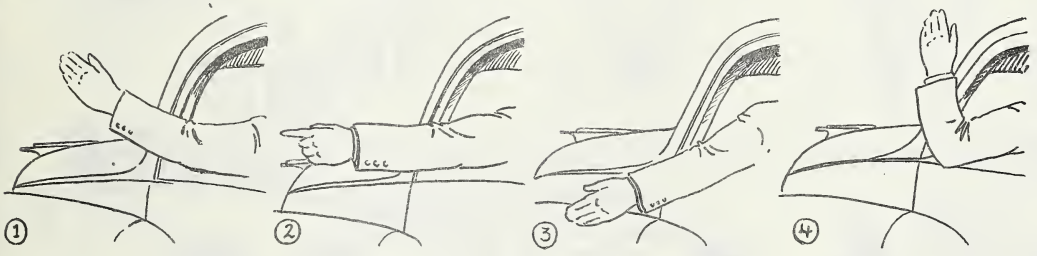
( )

67. Which one of these joints allows movement in all directions?



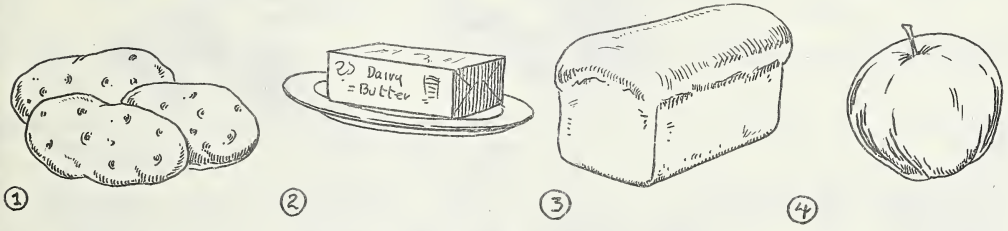
( )

. Which driver is about to make a right turn?



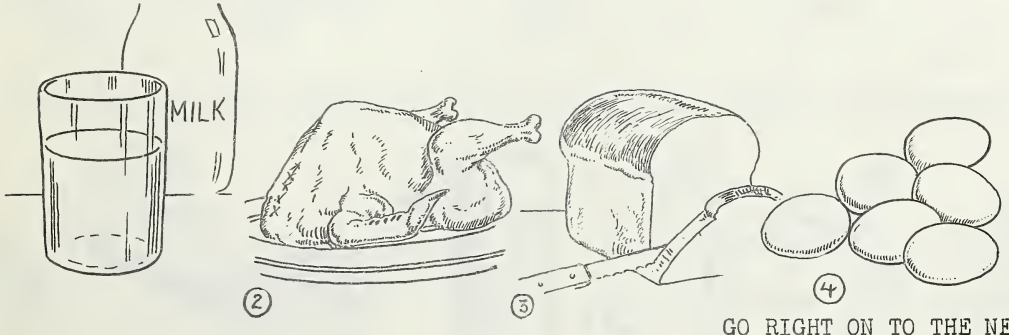
( )

69. Which type of food produces the most heat in the body?



( )

70. Which food has the least protein?

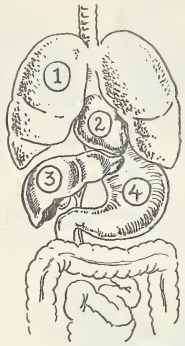


( )

GO RIGHT ON TO THE NEXT PAGE

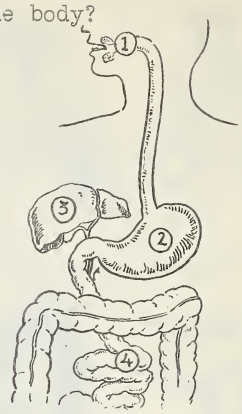
71. Which of these organs secretes a digestive juice and acts as a storehouse?

( )



72. Where do the digestive juices begin to work on food taken into the body?

( )



73. What is wrong with this picture

1. The man is putting on the light while washing.
2. The light has a chain switch.
3. The man hasn't finished dressing.
4. The basin is too full.

( )

74. What common cause of accidents is shown in the picture below? Write the number of the correct statement.

1. The truck is on the wrong side of the road.
2. One car should not try to pass another on a hill.
3. You should always keep to the right when passing a car.
4. The cars are travelling too fast.

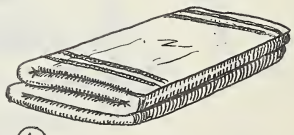
( )

75. A person's clothes have caught fire. What would you use to put out the fire?

①

②

③



④

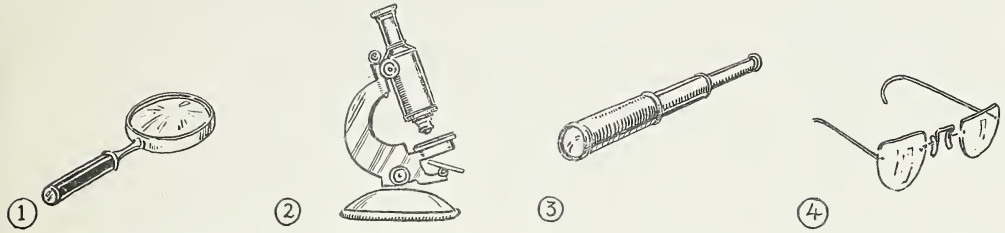
( )

GO RIGHT ON TO THE NEXT PAGE

76. A boy has discovered a decayed spot on one of his teeth. What should be done about it?

① ② ③ ④ ( )

77. Germs can be seen with one of the following. Which one would you use?

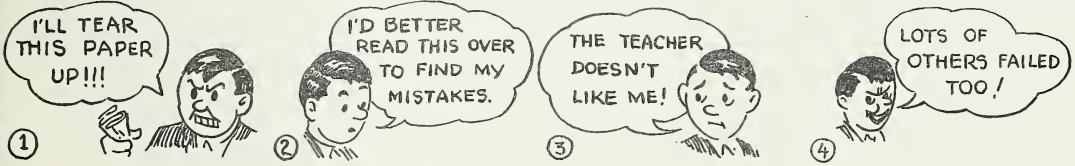


① ② ③ ④ ( )

78. We have several kinds of teeth which have different uses. Which one of these is used for grinding our food?

① ② ③ ④ ( )

79. You have just had an examination paper returned with a very low mark. What should your thoughts be?

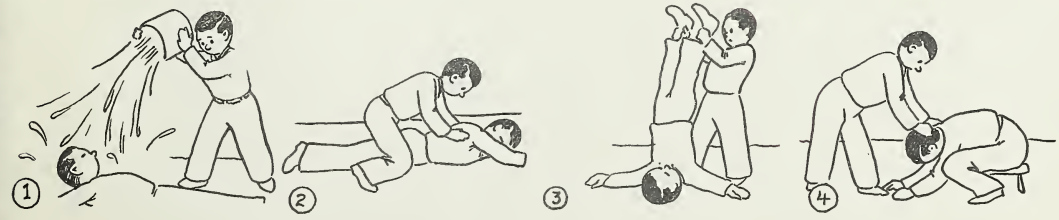


① ② ③ ④ ( )

80. If you haven't been able to get on the school hockey team what should you do?

① ② ③ ④ ( )

81. A person has been rendered unconscious in a fire. Which is the best way to revive him?



① ② ③ ④ ( )

B29763